UNIVERSIDAD DE LAS CIENCIAS INFORMÁTICAS
FACULTAD 10

Mecanismos Semiautomáticos y Automáticos para la Generación de Ontologías de Dominio.

Trabajo de Diploma para optar por el título de
Ingeniero en Ciencias Informáticas.

Autores:

Yolanda Sardiñas Suárez

César Miguel Calvo de la Paz

Tutores:

MSc. David Leyva Leyva

MSc. Daymy Tamayo Ávila

Lic. Vadim Fidel Turiño Sviatchenko

Ciudad de La Habana
Junio 2008

"Los filósofos se han limitado a interpretar el mundo de distintos modos; de lo que se trata es de transformarlo".

Karl Marx

Declaramos ser autores de este trabajo de diploma y concedemos a la Universidad de las Ciencias Informáticas los derechos patrimoniales de la misma para su beneficio.

Para que así conste firmamos la presente a los ____ días del mes de _____ del año 2008.

_____

Yolanda Sardiñas Suárez
Firma de la Autora

_____

César Miguel Calvo de la Paz
Firma del Autor

_____

MSc. David Leyva Leyva
Firma del Tutor

_____

MSc. Daymy Tamayo Ávila
Firma de la Tutora

_____

Lic. Vadim Fidel Turiño Sviatchenko
Firma del Tutor

MSc. David Leyva Leyva: Jefe de Departamento de Técnicas de Programación, Facultad 10, Universidad de las Ciencias Informáticas. Teléfono: (53) (07) 837 2511. davidl@uci.cu Licenciado en Cibernética-Matemática, Universidad de Las Villas (UCLV), 1988. Máster en Computación Aplicada, Universidad de Las Villas (UCLV), 1995.
Profesor de la Universidad de Holguín desde 1990. Jefe de Departamento de Informática (2003-2005). Participó en un Proyecto de Educación a Distancia en la Universidad del 2001 hasta el 2005, año en el que comenzó a trabajar en la Universidad de las Ciencias Informáticas. Ha participado en un gran número de eventos nacionales e internacionales. Le fueron concedidas sendas becas Intercampus en La Universidad de Castilla-La Mancha (1998) y La Universidad Autónoma de Madrid (2000). Trabajó como profesor invitado en la University of Belice (Belice, de 2001 a 2003) y de la Universidad Nacional de Ingeniería (Managua, Nicaragua, 2005).
Actualmente es líder del Polo Teleformación, grupo de investigación y desarrollo orientado a la extensión de la plataforma Moodle y al desarrollo y soporte de otras herramientas para la gestión de contenidos relacionados con el e-Learning.

MSc. Daymy Tamayo Ávila: Especialista de la Dirección de Teleformación, Profesora de Práctica Profesional V. Facultad 10, Universidad de las Ciencias Informáticas. Carretera a San Antonio de los Baños Km. 2 ½, Torrens, Boyeros, Ciudad de La Habana. Cuba. Teléfono: (53) (07) 837 2453. daymy@uci.cu Ingeniera Informática, Universidad de Holguín, 2005. Máster en Informática Aplicada, UCI-CUJAE, 2007.
Culminó sus estudios obteniendo Título de Oro. Fue Alumna de Alto Aprovechamiento Académico y Alumna Ayudante en la Disciplina de Técnicas de Programación de Computadoras desde el segundo año de su carrera. Participó en un Proyecto de Educación a Distancia en la Universidad del 2001 hasta el 2005, año en el que comenzó a trabajar en la Universidad de las Ciencias Informáticas. Ha participado en un gran número de eventos nacionales e internacionales. Ha sido tutora de varios Trabajos de Curso y Diploma, y de Alumnos Ayudantes. Es la arquitecta principal del Polo Teleformación, grupo de investigación y desarrollo orientado a la extensión de la plataforma Moodle y al desarrollo y soporte de otras herramientas para la gestión de contenidos relacionados con el e-Learning.

Lic. Vadim Fidel Turiño Sviatchenko: Profesor de Programación, Facultad 10, Universidad de las Ciencias Informáticas. Teléfono: (53) (07) 837 2543. vadim@uci.cu Licenciado en Cibernética-Matemática, Universidad de Las Villas (UCLV), 2007.
Desarrolló su Trabajo de Diploma en un tema relacionado con la Web Semántica, en específico Mapas Conceptuales.

Agradezco sobre todo a mis padres, por estar conmigo en cada paso que doy, por confiar en mí, por el ánimo, apoyo y alegría que me brindan, además de darme la fortaleza necesaria para seguir adelante.

A mi compañero César, por todo su cariño, comprensión, dedicación y paciencia que siempre ha mantenido conmigo. Por haber sido mi soporte y compañía durante todo el período de estudio.

Un agradecimiento especial a todos mis amigos, principalmente a Yadira, Benjamín, Henry y Alioscha, por compartir conmigo momentos tanto alegres como tristes, por tener siempre tendida su mano amiga, por escucharme, comprenderme y sobre todo por haberme brindado su amistad desde el día que me conocieron.

En general quisiera agradecer a todas y cada una de las personas que me han apoyado en la realización de esta tesis, con sus altos y bajos, compartiendo las angustias y las gratificaciones, a todos ellos gracias.

Yolanda Sardiñas Suárez

Agradezco en primer lugar a mis padres, por servirme de apoyo en cada decisión y cada paso que doy y por tenerme tanta confianza.

A mi compañera Yolanda por toda la comprensión que ha tenido conmigo y por haber terminado de educarme, ya que sin ella no hubiera podido ni siquiera pensar en estar aquí ahora. Para ella todo mi cariño, mi amor y mi agradecimiento.

También debo agradecer a todos mis amigos, como Henry, Benjamín, ya que con su amistad y sus críticas amigablemente fuertes han contribuido a mi superación.

César Miguel Calvo de la Paz

Dedico este trabajo a mis padres queridos, por su apoyo incondicional, por todo el cariño y toda una vida de constante dedicación y sacrificios.

A César, mi chachito, por brindarme su cariño, por tener tanta paciencia conmigo y ayudarme siempre.

A toda mi familia que de una forma u otra siempre me ha apoyado, se ha preocupado por mi bienestar, mi superación profesional. A todos ellos gracias.

Yolanda

Dedico este trabajo a mi familia, principalmente a Mayda y al Vlado, mis padres, por educarme y contribuir en gran parte al logro de éste éxito.

A Yolanda, mi chachita, por todo su cariño y por lo que me ha ayudado en este tiempo.

Especialmente dedico este trabajo a mi hermana para que lo use como guía y llegue más lejos en la vida.

César Miguel

En estos últimos años Internet se ha transformado en un gran repositorio de información ocupando un lugar significativo sobre todo para la esfera de la educación, y es debido a este considerable aumento de información y a la falta de organización de la misma, que los resultados arrojados por los motores de búsqueda no son los más satisfactorios. Para mejorar este proceso se ha propuesto gestionar el conocimiento prestando especial atención a las ontologías, transformando así la Web actual hacia una nueva Web denominada la Web Semántica. Esto ha traído consigo un aspecto negativo y es el trabajo engorroso que lleva crear las ontologías manualmente, ya que su número tiende a ser elevado a medida que se vaya incorporando información. Por tanto sería deseable que estas ontologías se generasen automáticamente o al menos semiautomáticamente.

Con el desarrollo de esta investigación se pretende contribuir al mejoramiento de la gestión del conocimiento, formulando una propuesta teórica para la creación de manera automática y semiautomática de las ontologías a partir de texto, para aplicarla en un futuro en entornos e-Learning, específicamente a los repositorios de objetos de aprendizaje, de manera tal que se conviertan en repositorios semánticos.

Para la realización de este trabajo se hizo un profundo estudio sobre los aspectos teóricos relacionados con el tema de investigación, abordando algunas definiciones de las ontologías, las principales características de herramientas y lenguajes empleados en la generación de las mismas, así como un breve análisis sobre el dominio destinado para la aplicación de esta propuesta. Luego se llegó a la elaboración de una teoría para generar las ontologías automáticamente y semiautomáticamente que consta de tres fases, en cada una de las cuales se presentaron varias herramientas que intervienen en este proceso. Finalmente se explicó con tres ejemplos cómo sería el funcionamiento de un sistema que pusiera en práctica la propuesta teórica planteada en esta investigación.

Palabras clave: Gestión del Conocimiento, Web Semántica, Ontologías, e-Learning, Metadatos, Objetos de Aprendizaje.

Índice de Contenido

Índice de Figuras

Índice de Tablas

Introducción

Actualmente Internet se ha convertido en objeto de atención y trabajo, debido a la gran cantidad de información y servicios que se puede encontrar en ella, también por lo útil que resulta su uso en algunas esferas como la medicina, la educación, las investigaciones científicas, entre otras. Sin embargo, esta red mundial presenta además de sus enormes beneficios, algunos inconvenientes que hacen que su éxito esté incompleto, y es la dificultad con que realiza la búsqueda de información debido a la falta de organización de sus contenidos, provocando en ocasiones que se obtengan resultados en la búsqueda, que no guarden relación con el contenido deseado. Esto ocurre mayormente por la saturación de información que existe en la red global. En Internet todos publican lo que desean, pero no clasifican su información, es decir, no todos los autores usan métodos distintivos y/o clasificatorios para determinar tipo y tema de los datos publicados. No obstante, a pesar de los importantes avances aportados por las nuevas tecnologías, el usuario de la Web aún carece de un sistema que permita procesar y acceder a la información documental contenida en sitios Web de una manera fiable.

Hoy día la búsqueda de información a través de algunos buscadores Web, es procesada recuperando aquellos documentos en los cuales aparecen los elementos constituyentes de la consulta del usuario, ya que este modo de búsqueda localiza los términos de la consulta tal y como se han introducido. Esto puede traer consigo una serie de errores como los citados a continuación:

Polisemia: Se realizan consultas de un término. Los resultados devueltos incluyen documentos con el término solicitado, pero con un significado distinto.

Sinonimia: Se realiza una consulta que devuelve muchos documentos que en ocasiones no son los correctos, esto es debido a que en los textos aparece un sinónimo del término de la consulta y no el propio término.

Multilingüismo: Se realiza una consulta con términos en un determinado idioma que no está presente en los textos. Sin embargo, el término está presente en otro lenguaje.

Estos problemas se deben a que el motor de búsqueda no identifica de forma precisa el contenido en el que se está interesado, puesto que al no hacer inferencia sobre la consulta no recupera textos relevantes.

En los últimos años, muchos investigadores están diseñando modelos para transformar la red de un espacio de información a un espacio de conocimientos. Recientemente, Tim Berners-Lee, uno de los pioneros de la Web y director del W3C[1], defiende el desarrollo de la Web con conocimiento. Esta red dotada de una estructura que permitiera expresar el contenido de las páginas de una forma que los ordenadores pudieran "entenderlas" y que posibilitase tanto la interacción entre ordenadores como entre éstos y los usuarios, proponía así un nuevo modelo en el que todos sus contenidos estarían descritos y estructurados de un modo que las máquinas podrían comprenderlos, permitiría el procesamiento de la información y de los metadatos[2], los cuales describirían la información disponible en la red, para de esta forma poder lograr una mejor gestión del conocimiento[3].

Debido al problema de la no clasificación de la información, que también está vigente en la Universidad de las Ciencias Informáticas (UCI), el trabajo de diploma realizado por (Cabrera & Martínez, 2007) se basa en la propuesta de una metodología que gestiona el conocimiento mediante ontologías[4], donde plantean de manera general que la solución más óptima para resolver el problema que presenta la Web actual era la Web Semántica[5]. Esta propuesta metodológica fue creada con el propósito de aplicarla en un futuro a los repositorios[6], en este caso, de objetos de aprendizaje, donde los recursos contenidos en éstos estuvieran organizados y estructurados, y sobre todo relacionados entre ellos según el tema sobre el que tratasen. Pero para ello era necesario que la información guardada en estos orígenes de datos, se encontrara estructurada mediante ontologías, de forma tal que facilitara el razonamiento automatizado del mismo, donde los usuarios pudieran especificar sus

[1] ***W3C**: World Wide Web Consortium*

[2] ***Metadatos:** "Conjunto de atributos o elementos necesarios para describir un recurso en cuestión" (López, 2006).*

[3] ***Gestión del Conocimiento:** "Proceso sistemático de buscar, organizar, filtrar y presentar la información con el objetivo de mejorar la comprensión de las personas en un área específica de interés" (Figueroa, 2006).*

[4] ***Ontologías:** "Estructura jerárquica que define formalmente las relaciones semánticas de un conjunto de conceptos. Se usa para crear vocabularios estructurados para la recuperación o el intercambio de información" (Woodley, 2003).*

[5] ***Web Semántica:** "Extensión de la Web actual que cuenta con información debidamente estructurada, lo que proporciona un significado bien definido. Mejora la forma en la que las máquinas y las personas trabajan en cooperación" (Samper, 2005).*

[6] ***Repositorio:** "Sitio centralizado donde se almacena y mantiene información digital, habitualmente bases de datos y archivos informáticos" (Wikipedia, 2008).*

criterios de búsquedas en función de conceptos y atributos que se establecerían inteligentemente. Su utilidad se basa en poder expandir la consulta realizada por el usuario. Navegando por la estructura jerárquica que define una ontología es posible recuperar los términos que tengan relación con las palabras clave de la petición. De esta manera los datos buscados en las colecciones de documentos, son más concretos, precisos y completos.

El trabajo desarrollado por (Cabrera & Martínez, 2007) daba cumplimiento a los objetivos que en él se trazaban, ya que, con la propuesta desarrollada, basada en una serie de pasos para crear ontologías manualmente, se evitaban los inconvenientes citados en la búsqueda de información al lograr una estructuración de la misma.

Crear las ontologías de forma manual trae consigo un trabajo muy engorroso para el personal informático que lleve a cabo la realización de esta tarea, ya que su número tiende a ser elevado a medida que se vaya incorporando información en los repositorios.

Con ánimo de continuar y perfeccionar el trabajo realizado por (Cabrera & Martínez, 2007), se propone construir las bases teóricas para la generación automática o semiautomática de ontologías de dominio, que facilite la fusión de la Web Semántica y los repositorios, mediante la futura implementación de un repositorio semántico que aplique la teoría propuesta en esta investigación.

Problema a resolver: *¿Cómo favorecer los procesos de gestión del conocimiento a través de mecanismos de generación de ontologías?*

Para solucionar este problema el presente trabajo tiene como **objetivo general** *la confección de una propuesta metodológica para la generación automática o semiautomática de ontologías de dominio.*

Se define como **objeto de estudio** *la gestión del conocimiento basada en ontologías* y enmarcando como **campo de acción** *la generación automática o semiautomáticamente de ontologías de dominio*.

Para el desarrollo de la investigación se proponen las siguientes preguntas científicas:

1. ¿Cuáles son los aspectos teóricos que sustentan el proceso de generación automática o semiautomática de ontologías?

2. ¿Existirán herramientas efectivas para la generación de ontologías de manera automática o semiautomática?
3. ¿Se podrá elaborar una propuesta teórica eficiente para la generación automática o semiautomática de ontologías de dominio para la gestión del conocimiento?

Con el objetivo de dar respuesta a las preguntas anteriores se definieron las siguientes tareas:

1. Analizar los diferentes conceptos relacionados con el tema de la generación automática o semiautomática de ontologías.
2. Realizar un estudio de las herramientas utilizadas para la generación de ontologías de dominio de modo semiautomático y automático.
3. Realizar una propuesta teórica para la vinculación de mecanismos de generación de ontologías de dominio con la gestión del conocimiento.
4. Caracterizar herramientas que intervienen en la propuesta teórica para la generación de ontologías.
5. Realizar un estudio sobre los aspectos teóricos generales relacionados con el e-Learning, específicamente los ROA[7], dominio en el que será aplicada la propuesta.

Para darle cumplimiento a las tareas de investigación se utilizaron varios métodos teóricos y empíricos:

Dentro de los métodos teóricos se utilizaron:

El **Analítico-Sintético** para encontrar la relación entre los diferentes conceptos involucrados en los mecanismos automáticos y semiautomáticos para la generación de ontologías de dominio, y para realizar una propuesta de solución basada en el estudio de esos mecanismos que sea efectiva para la gestión del conocimiento.

La **Modelación** para realizar un estudio de las diferentes herramientas utilizadas en el proceso de generación automática y semiautomática de ontologías de dominio, sus cualidades y

[7] ***ROA:*** *Repositorio de Objetos de Aprendizaje*

peculiaridades, para lograr el correcto desarrollo de una propuesta para la automatización de ontologías, que permitiera predecir la ontología creada por dicho proceso a variaciones de algunos de sus parámetros sin tener que ejecutar el proceso en la realidad.

Y como métodos empíricos se utilizaron:

La **Observación Sistémica** para poder conocer el funcionamiento de herramientas software para la generación de ontologías y emitir una valoración sobre ellas, para realizar este proceso genérico o plantear uno nuevo y mejor mecanismo de generación.

La **Revisión de Documentos** como una técnica subordinada al método de análisis y síntesis para comprender los conceptos fundamentales y el funcionamiento de las distintas herramientas documentadas en la bibliografía consultada y la forma en que éstas operan.

El presente trabajo consta de introducción, tres capítulos, conclusiones generales, recomendaciones, referencias bibliográficas y bibliografía utilizadas durante el desarrollo del trabajo, un glosario de términos y por último, los anexos que complementan el cuerpo del trabajo.

Capítulo 1: Fundamentación Teórica. Se analizan los principales conceptos relacionados con el tema de investigación, así como las características fundamentales de herramientas y lenguajes existentes para la generación automática o semiautomática de ontologías. Además, se muestran varias formas de representar las ontologías y se realiza un estudio sobre los aspectos teóricos relacionados con el e-Learning, específicamente los ROA, dominio en el que será aplicada la propuesta.

Capítulo 2: Propuesta Metodológica. Se llega a la formulación de una teoría para regir la creación de un sistema de generación automática o semiautomática de ontologías de dominio. Además, se presentan herramientas que intervienen en la propuesta teórica y se emiten valoraciones sobre ellas.

Capítulo 3: Análisis de Ejemplos. Se realiza un análisis teórico mediante ejemplos que muestran cómo sería la aplicación de la propuesta que se presenta en el Capítulo 2, para crear las ontologías automáticamente y semiautomáticamente.

Capítulo 1: Fundamentación Teórica

1.1. Introducción

En el presente capítulo se brinda una visión general de algunos conceptos relacionados con las ontologías, necesarios para comprender el tema que se aborda en este trabajo de investigación. Se muestran además, los distintos tipos de ontologías que existen, algunos lenguajes y herramientas utilizadas en la generación automática o semiautomática de ontologías, así como los beneficios que reporta el aprendizaje de ontologías. También se presentan distintas formas de representar las ontologías y se hace referencia a los aspectos teóricos relacionados con el e-Learning y los repositorios, como objetos de aprendizaje y metadatos, dominio de aplicación en que estará enmarcada la propuesta teórica para la generación automática o semiautomática de ontologías.

1.2. Ontologías

1.2.1. Definiciones

En los últimos años, la creación de software basado en ontologías está aumentando considerablemente. El término ontología, actualmente, constituye un tópico de interés por parte de toda la comunidad científica, independientemente del ámbito de investigación. Su uso como herramienta de categorización de información en la Web es uno de los más conocidos.

Existe un grupo de definiciones desde diferentes puntos de vista acerca de este término.

1.2.1.1. Definiciones relacionadas con la Filosofía

En las siguientes líneas se presentan diferentes definiciones propuestas por importantes filósofos.

“Ontología es la ciencia de algo y de nada, del ser y del no ser, de la cosa y del modo de la cosa, de la sustancia y el accidente” (Leibniz) (Couturat, 1903).

"La filosofía trascendental es el sistema de todas nuestras cogniciones puras a priori, que podemos llamar ontología. Así, ontología trata con cosas en general, desde abstractas hasta particulares. Abarca todos los conceptos puros de la comprensión y todos los principios de la razón. Las ciencias principales que pertenecen a la metafísica son: ontología, cosmología, y teología. Ontología es una pura doctrina de elemento de toda nuestra cognición al completo, o: contiene la suma de todos nuestros conceptos puros que podemos tener a priori sobre la cosas" (Kant, 2001).

"La gente trata con asuntos relacionados con la teoría de entidades desde la antigüedad bajo el título de 'Metafísica' y, especialmente, bajo el título de 'Ontología' como parte de la metafísica; y ellos no han fallado siempre a reconocer las características de la libertad de existencia" (Meinong, 1921).

La siguiente definición fue tomada del diccionario "Webster's Third New Internacional Dictionary" (*WTNID, 2008*):

"Ciencia o estudio del ser: específicamente, una rama de la metafísica relacionada con la naturaleza y las relaciones del ser; un sistema particular según el cual se investigan los problemas de la naturaleza del ser; esto es, filosofía fundamental".

1.2.1.2. Definiciones relacionadas con Inteligencia Artificial

Una antigua definición de ontología en Inteligencia Artificial apareció en *(Neches et al, 1991):*

"Una ontología define los términos básicos y relaciones que conforman el vocabulario de un área específica, así como las reglas para combinar dichos términos y las relaciones para definir extensiones de vocabularios".

Una de las definiciones más extendidas es la dada por Tom Gruber *(Gruber, 1993):*

"Una ontología es una especificación explícita de una conceptualización. El término proviene de la filosofía, donde una ontología es un recuento sistemático de la existencia. En sistemas de Inteligencia Artificial, lo que existe es lo que puede ser representado. Cuando el conocimiento de un dominio se representa mediante un formalismo declarativo, el conjunto de objetos que puede ser representado se

llama universo del discurso. Esos conjuntos de objetos, y las relaciones que se establecen entre ellos, son reflejados en un vocabulario con el cual representamos el conocimiento en un sistema basado en conocimiento. Así, en el contexto de IA, podemos describir la ontología de un programa como un conjunto de términos. En tal ontología, las definiciones asocian nombres de entidades del universo del discurso con textos comprensibles por los humanos que describen el significado de los nombres, y axiomas formales que limitan la interpretación y buen uso de dichos términos. Formalmente, una ontología es una teoría lógica".

Otra definición de ontología es la presentada por Borst *(Borst, 1997):*

"Una ontología es una especificación formal de una conceptualización compartida."

Posteriormente, las definiciones de Gruber y Borst fueron explicadas en *(Studer et al, 1998)* de la siguiente forma:

"Conceptualización se refiere a un modelo abstracto de algún fenómeno en el mundo a través de la identificación de los conceptos relevantes de dicho fenómeno. Explícita significa que el tipo de conceptos y restricciones usados se definen explícitamente. Formal representa el hecho de que la ontología debería ser entendible por las máquinas. Compartida refleja la noción de que una ontología captura conocimiento consensual, esto es, que no es de un individuo, sino que es aceptado por un grupo."

Otra definición tomada del diccionario "Webster's Third New Internacional Dictionary" (*WTNID, 2008*):

"Teoría relativa a los tipos de entidades y específicamente los tipos de entidades abstractas que se admiten en el lenguaje de un sistema".

La siguiente definición es la que más se ajusta al tema de investigación de este trabajo:

"Estructura jerárquica que define formalmente las relaciones semánticas de un conjunto de conceptos. Se usa para crear vocabularios estructurados para la recuperación o el intercambio de información" (Woodley, 2003).

1.2.2. Objetivos

Según lo expuesto en *(Guzmán, 2006)*, las ontologías tienen como objetivos:

- Compartir la comprensión común de la estructura de información entre personas o agentes de software.
- Permitir la reutilización del conocimiento perteneciente a un dominio.
- Permitir representar de manera explícita los supuestos de un dominio.
- Separar el conocimiento de un dominio, del conocimiento que se pueda denominar operacional.
- Permitir analizar el conocimiento de un campo.

1.2.3. Tipos de Ontologías

En *(Carrera, 2007)* se propone la siguiente clasificación de ontologías, teniendo en cuenta diferentes criterios:

Alcance de su aplicabilidad

- Ontologías de dominio: Proporcionan un vocabulario necesario para describir un dominio dado. Incluyen términos relacionados con los objetos del dominio y sus componentes.
- Ontologías de tarea: Proporcionan un vocabulario para describir términos involucrados en los procesos de resolución de problemas, los cuales pueden estar relacionados con tareas similares en el mismo dominio o en dominios distintos. Incluyen nombres, verbos, frases y adjetivos relacionados con la tarea.
- Ontologías generales: Representan los datos generales que no son específicos de un dominio. Son las ontologías de nivel más alto ya que describen conceptos generales (espacio, tiempo, materia, objeto, etc.).
- Ontologías específicas: Son ontologías especializadas que describen los conceptos para un campo limitado del conocimiento o una aplicación concreta.

Granularidad de la conceptualización

- Ontologías terminológicas: Especifican términos que son usados para representar conocimiento en el universo de discurso. Suelen ser usadas para unificar vocabulario en un dominio determinado.
- Ontologías de información: Ofrecen un marco para el almacenamiento estandarizado de información.
- Ontologías de modelado del conocimiento: Especifican conceptualizaciones del conocimiento. Poseen una rica estructura interna y suelen estar ajustadas al uso particular del conocimiento que describen (términos y semántica).

Agentes que vayan a emplear la ontología

- Ontologías lingüísticas: Se vinculan a aspectos lingüísticos tales como los gramaticales, semánticos y sintácticos, destinados a su utilización por los seres humanos.
- Ontologías no lingüísticas: Aquellas ontologías destinadas a ser empleadas por robots o agentes inteligentes.
- Ontologías mixtas: Combinan las características de las anteriores.

Nivel de abstracción y razonamiento lógico permitido

- Ontologías descriptivas: Estas ontologías incluyen descripciones, taxonomías de conceptos, relaciones entre los conceptos y propiedades, pero no permiten inferencias lógicas.
- Ontologías lógicas: Estas ontologías permiten inferencias lógicas mediante la utilización de una serie de componentes como la inclusión de axiomas, etc.

Entre las ontologías de mayor utilidad en la gestión del conocimiento están presentes las ontologías de dominio, encargadas de modelar el conjunto de conceptos de un dominio y las relaciones que existen entre ellos, según como lo entiende una comunidad científica. Es por esto que se propone el uso de mecanismos automáticos y semiautomáticos para la generación de ontologías de dominio, a fin de representar las relaciones entre los conceptos de las diversas áreas del conocimiento a las que

pertenecen los OA[8] de un repositorio. Estas ontologías pueden, además, ser de naturaleza multilingüe ya que contribuye a una mejor representación, búsqueda y recuperación de los OA, lo que se convierte en una fortaleza dentro del repositorio, sobre todo en entornos e-Learning, donde la información es más importante que el idioma de representación.

1.3. Ingeniería Ontológica

Con el fin de mitigar el problema de la elaboración manual de ontologías que supone un coste tan elevado en tiempo y dinero, y con la meta futura de la creación y puesta en marcha de la Web Semántica, se ha creado una nueva disciplina, la ingeniería ontológica, dedicada al estudio y diseño de entornos y aplicaciones que ayuden a elaborar, mantener y utilizar ontologías (Pedraza, 2007).

La ingeniería ontológica comprende por tanto, el conjunto de actividades concernientes al proceso de desarrollo de las ontologías, a su ciclo de vida, los métodos y metodologías para construirlas y las herramientas y lenguajes que las soportan *(Gómez).*

La ingeniería ontológica surgida de la Web Semántica, proporciona los requisitos necesarios para mejorar las búsquedas de información. Esta mejora se debe a que en lugar de utilizar palabras clave en los procesos de búsqueda, se centra en el significado de los conceptos, es decir, en la semántica de la información. De esta manera, se obviará que los datos deben ser entendidos exclusivamente por los usuarios y se pasará a un proceso de entendimiento recíproco entre el hombre y la máquina, en el que las máquinas pasarán a "comprender" los datos que procesan, actuando sin la necesaria y continuada supervisión actual.

Según lo expuesto en *(Fillottrani, 2007)* la ingeniería ontológica se basa en los siguientes principios:

- No existe una única forma de modelar un dominio, siempre existen alternativas viables. La mejor solución siempre depende de las aplicaciones, y sus futuras extensiones.

[8] ***OA*** *(Objetos de Aprendizaje): "Entidades digitales con características de diseño instruccional, que pueden ser usadas, reutilizadas o referenciadas durante el aprendizaje soportado en computadora con el objetivo de generar conocimientos, habilidades y actitudes en función de las necesidades del alumno" (Galeana).*

- Necesariamente el desarrollo de una ontología es un proceso iterativo.
- Los conceptos deben ser cercanos a objetos físicos o lógicos, y las relaciones deben ser aquellas interesantes para el dominio.

1.3.1. Características de un entorno para la Ingeniería Ontológica

Un sistema para la generación automatizada de ontologías debe *(Mizoguchi, 2004)*:

- Permitir la gestión de todo el proceso de desarrollo de la ontología.
- Facilitar el desarrollo colaborativo.
- Poseer una metodología fundamentada en la teoría de ontologías.
- Representar formalmente la ontología resultante mediante alguna norma o sintaxis (normalmente las recomendaciones del W3C).
- Disponer de un motor de inferencias.
- Ser usable.
- Ser extensible.

1.3.2. Lenguajes empleados en la Construcción de Ontologías

1.3.2.1. Lenguaje XML (Extensible Markup Language)

XML es un lenguaje para estructurar la información en un documento o en general, en cualquier fichero que contenga texto. Además, este formato es usado para la gestión de Web Services[9] y ha ganado muchísima popularidad en los últimos años por ser un estándar abierto y libre, creado por el W3C.

Para un dominio particular, XML es una buena alternativa como lenguaje de marcado semántico, utilizando esquemas XML para definir vocabularios, o bien una combinación de XML y RDF. Se

[9] *Un **Servicio Web** (en inglés Web Service) es una colección de protocolos y estándares, cuya función es el intercambio de datos entre aplicaciones a través de Internet. La interacción de aplicaciones de distintos sistemas operativos permite que programas de muy diversa concepción se combinen y proporcionen servicios integrados. Ésta es la nueva Internet, centrada en las aplicaciones y basada en los servicios (Banespyme).*

presentan a continuación algunas características de XML que se han tenido en cuenta a la hora de elegirlo como formato predeterminado en la elaboración de las ontologías *(Carrera, 2007).*

- Es un estándar basado en un conjunto de reglas para la definición de etiquetas semánticas que organizan los documentos en diferentes secciones.
- Es una arquitectura más abierta y extensible. Los identificadores pueden crearse de manera sencilla y ser adaptados.
- Mayor consistencia, homogeneidad y amplitud de los identificadores descriptivos de los documentos.
- La codificación del contenido en XML consigue que la estructura de la información resulte más accesible. Además, existe independencia entre el contenido de los datos y la presentación de los mismos.
- Es un formato ideal para guardar datos de configuración de las aplicaciones.

1.3.2.2. Lenguaje XML Schema

XML Schema o Esquema XML, es un lenguaje de esquema utilizado para describir la estructura y las restricciones de los contenidos de los documentos XML de una forma muy precisa, más allá de las normas sintácticas impuestas por el propio lenguaje XML. Se consigue así, una percepción del tipo de documento con un nivel alto de abstracción. Fue desarrollado por el W3C y alcanzó el nivel de recomendación en mayo de 2001 *(Wikipedia, 2008).*

El principal aporte de XML Schema es el gran número de tipos de datos que incorpora. De esta manera, XML Schema aumenta las posibilidades y funcionalidades de aplicaciones de procesado de datos, incluyendo tipos de datos complejos como fechas, números y strings.

1.3.2.3. Lenguaje RDF (Resource Description Framework)

RDF es un lenguaje creado por el W3C y está orientado a la creación de metadatos para la descripción de recursos Web de forma distribuida e interoperable. Está muy relacionado con XML, pues RDF hace

posible la especificación de información semántica mediante la utilización de XML. Proporciona un modelo de datos simple para describir la semántica de la información de una forma accesible por la máquina. Es un modelo para representar propiedades etiquetadas y valores de propiedades que describen a los objetos (Codina, 2007).

1.3.2.4. Lenguaje RDF Schema

RDF Schema o Esquema RDF**,** es una extensión semántica de RDF. Un lenguaje primitivo de ontologías que proporciona los elementos básicos para la descripción de vocabularios. La versión más reciente fue publicada en febrero de 2004 también por el W3C.

Características principales (*Wikipedia, 2008)*:

- Un archivo RDFS es un archivo RDF, es decir, se trata de un archivo con la misma sintaxis y la misma estructura que la que se usa en RDF. La sintaxis está basada en XML.
- Es extensible, cada desarrollador puede extender el esquema RDF de manera independiente.

1.3.2.5. Lenguaje OWL (Web Ontology Language)

OWL es un lenguaje de etiquetado semántico para publicar y compartir ontologías en la Web. Se trata de una recomendación del W3C, y puede usarse para representar ontologías de forma explícita, es decir, permite definir el significado de términos en vocabularios y las relaciones entre aquellos términos *(Cabrera & Martínez, 2007).*

OWL extiende de RDF y mejora algunas de sus limitaciones *(Lovillo, 2007):*

- Incluye propiedades que permiten restringir las instancias de una clase.
- Permite restringir los valores de una clase.
- Facilita expresar ciertas propiedades de las clases.
- Se puede especificar el número de elementos que participan en una relación.

OWL fue creado pensando en su flexibilidad. Este hecho viene dado debido a la naturaleza distribuida y dinámica de la Web. Esto, permite además, extender ontologías o emplear diversas ontologías ya existentes para completar la definición de una nueva ontología.

1.3.3. Aprendizaje de Ontologías

La disciplina conocida como "Aprendizaje de Ontologías", es una parte de la ingeniería ontológica que investiga el desarrollo de métodos y herramientas que permitan la creación de una ontología de forma semiautomática. Concretamente, se centra en la generación de herramientas que permitan importar, extraer, podar, refinar y evaluar la taxonomía[10] de una ontología. Estas aplicaciones estarán al servicio de un experto humano, el "ingeniero ontológico", que supervisará todo el proceso de creación de las ontologías (Pedraza, 2007).

En este proceso se emplea texto, diccionarios electrónicos, ontologías lingüísticas e información estructurada y semiestructurada para extraer conocimiento. Recientemente, gracias al enorme crecimiento de la sociedad de la información, la Web se ha convertido en una valiosa fuente de información para casi cualquier dominio. Esto ha provocado que los investigadores empiecen a considerar a la Web como un repositorio válido para recuperar información y adquirir conocimiento *(Sánchez, 2007).*

1.3.3.1. Arquitectura de sistemas para el Aprendizaje de Ontologías

A continuación, para ilustrar el funcionamiento de estos sistemas se describe brevemente la arquitectura de una de las primeras propuestas formuladas en este ámbito, la de Maedche (2001), que ha determinado las líneas básicas a seguir en este campo. La arquitectura propuesta consta de cuatro elementos:

[10] *Una **taxonomía** es un conjunto organizado de términos utilizado para organizar información y pensado principalmente para poder navegar por él. En otras palabras, una taxonomía simplemente exige que sus componentes estén organizados de manera que se puedan "recorrer". Su rasgo fundamental y definitorio es, por tanto, su finalidad de exploración. Son usadas, por ejemplo, para la categorización y la organización de la información en sitios Web (Centelles, 2005).*

- Interfaz gráfica: permite al ingeniero ontológico intervenir manualmente en todo el proceso de creación.

- Componente de gestión: con ella se seleccionan los datos a partir de los cuales construir la ontología (documentos HTML y XML, DTDS, bases de datos, otras ontologías, etc.).

- Centro de procesamiento de recursos: facilita al ingeniero ontológico diferentes herramientas para procesar los documentos de entrada y extraer la terminología necesaria (conceptos).

- Por último, estos sistemas deben disponer de una biblioteca de algoritmos cuyo funcionamiento se basa normalmente en reglas de asociación, técnicas de análisis formal de conceptos, o técnicas de agrupamiento (jerárquicas o no). Mediante la aplicación de uno o varios de estos algoritmos sobre los documentos ya procesados podrán extraerse las clases de la taxonomía y sus relaciones.

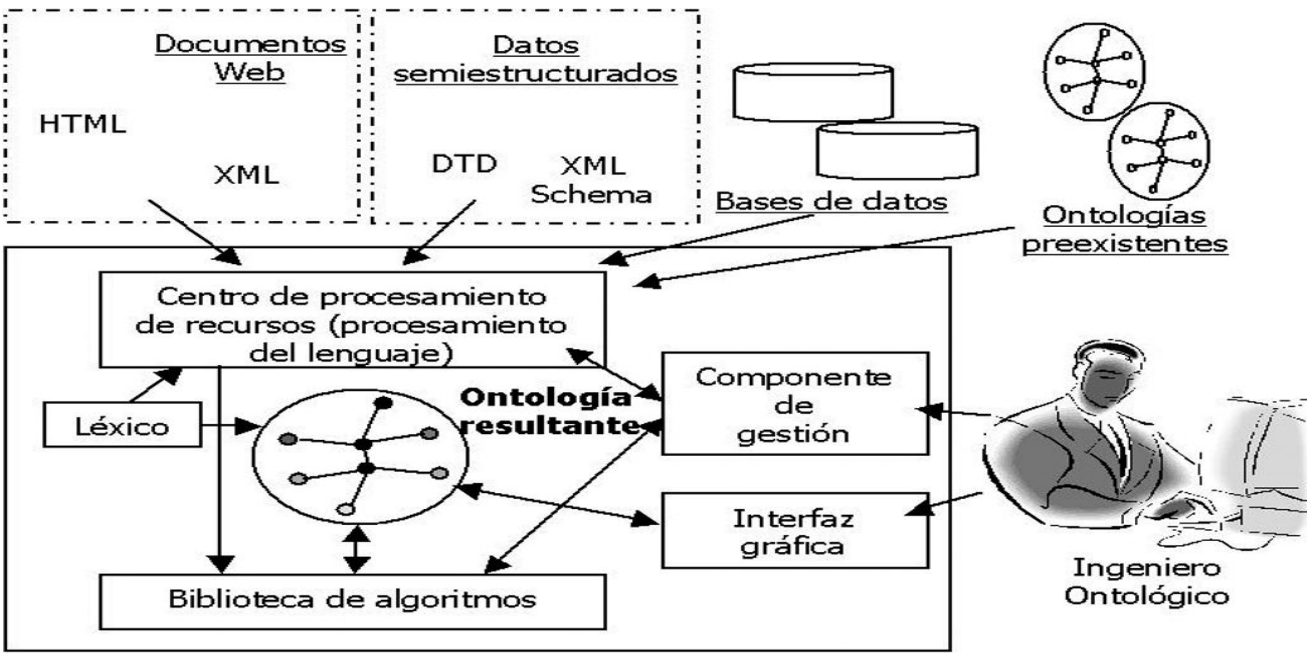

Figura1: Arquitectura del sistema para el aprendizaje de ontologías propuesto por Maedche (2001)

1.3.3.2. Etapas

Mediante el uso de sistemas como los descritos, podría construirse una ontología siguiendo las siguientes etapas *(Pedraza, 2007)*:

1. Importación y reutilización

Su objetivo es desarrollar mecanismos y estrategias para importar y reutilizar conceptualizaciones de un dominio a partir de estructuras o esquemas preexistentes.

- Identificación de las estructuras o esquemas, y discusión con los expertos en el dominio.
- Fusión de los esquemas y estructuras seleccionadas para constituir una única base sobre la cual elaborar la ontología y a partir de la cual aplicar el resto de las fases.

2. Extracción de la ontología

- La extracción de ontologías conlleva las siguientes fases:

1. Selección de los conceptos del dominio y sus léxicas (términos), que se obtienen en función del procesamiento de los textos de los recursos del dominio seleccionados (documentos HTML, etc.). Independientemente de la recomendación hecha por el sistema, el ingeniero ontológico puede incluir o eliminar entradas léxicas si así lo desea.
2. Generación de la taxonomía de conceptos: mediante técnicas de agrupamiento.

- Estas dos fases principales pueden ser complementadas con:

3. Ampliación de las relaciones entre conceptos mediante diccionarios y reglas de asociación.

El resultado final de esta fase es la propuesta de una taxonomía del dominio al ingeniero ontológico que éste puede modificar o rehacer como crea conveniente.

3. Poda

- El sistema debe permitir al ingeniero ontológico ajustar la ontología a su propósito original, eliminando un concepto cuando éste lo estime conveniente, pero mostrando las consecuencias de tal eliminación sobre el resto de la ontología.
- El sistema debe intentar mostrar la importancia de los conceptos dentro de la ontología, para que el ingeniero ontológico decida si debe mantenerlos o eliminarlos.

4. Refinamiento

El sistema pone a disposición del profesional herramientas que permiten completar y afinar el resultado final, incorporando nuevas entradas léxicas y/o conceptos como consecuencia de:

- Las necesidades específicas de los usuarios de un dominio.
- La actualización del dominio.

5. Evaluación de la ontología resultante

- A través del seguimiento y observación de su uso.

1.3.3.3. Beneficios de la Generación Semiautomática y Automática de Ontologías

La ingeniería ontológica es muy ventajosa para gestionar el conocimiento, pero esas ventajas pueden aumentar considerablemente si las ontologías se generasen de manera automática o semiautomática, proceso mediante el cual se alcanzan los siguientes beneficios:

- Se reduce enormemente el tiempo empleado en generar las ontologías, ya que ahorra un gran trabajo al experto en el dominio encargado de crear las ontologías manualmente.

- Disminuye considerablemente los enormes costos necesarios para crear las ontologías.

- Se utiliza cuando el dominio o la tarea en cuestión suponga una tarea inabarcable para el experto humano, bien por su magnitud, bien por su naturaleza cambiante.

- Proporciona una forma potente y eficiente de representar y compartir el conocimiento dentro de un área, utilizando para ello un vocabulario común.

- No requiere un gran esfuerzo intelectual.

1.4. Herramientas para el Aprendizaje de Ontologías

Las herramientas para el aprendizaje de ontologías se centran en el estudio de grandes corpus[11] normalmente en torno a un dominio concreto, para crear ontologías de manera automática mediante el uso de técnicas de Procesamiento de Lenguaje Natural. Dependiendo de la aplicación de la ontología, estas herramientas pueden requerir una mayor o menor supervisión, ya que en muchos casos los resultados pueden ser muy variados en función del corpus utilizado. Algunas de ellas se encuentran integradas en los propios editores, como es el caso de Text-to-Onto en Protégé[12].

1.4.1. Text-to-Onto

En (Maedche & Staab, 2001), se presenta una arquitectura de aprendizaje de ontologías para la Web Semántica. Esta arquitectura incluye un sistema para la construcción de ontologías a partir de texto denominado Text-to-Onto (Maedche & Staab, 2000), el cual ha sido desarrollado por el Instituto AIFB de la Universidad de Karlsruhe. Este sistema utiliza un algoritmo de aprendizaje de reglas de asociación (association-rule-learning) para descubrir relaciones entre conceptos dejando la decisión final de la creación de los nuevos conceptos y la inserción de las relaciones al ingeniero de conocimiento. El sistema utiliza un algoritmo para descubrir reglas de asociación generalizadas. Los datos de entrada están formados por frases y el conjunto de conceptos que aparecen juntos en la

[11] *Se refiere a un texto.*
[12] ***Protégé:*** *editor de ontologías que permite construir, modificar y consultar ontologías*.

frase. El algoritmo extrae las reglas de asociación representadas por conjuntos de elementos que aparecen juntos a menudo y le presenta las reglas al ingeniero de conocimiento.

El resultado del proceso de aprendizaje es una ontología de dominio que contiene conceptos de un dominio específico y conceptos independientes a un dominio *(Gómez, 2003).*

Meta de la herramienta: Encontrar relaciones taxonómicas y no taxonómicas.
Técnicas de aprendizaje usadas por la herramienta: Aproximación estadística, reglas de asociación y técnicas de poda.
Método usado para el aprendizaje de ontología: Método basado en el algoritmo Srikant y Agrawal (Srikant & Agrawal, 1995).
Intervención del usuario/experto en el proceso: Validación.
Tipos de recursos usados por el método: Diccionarios legibles por computadora y otras ontologías.
Interoperabilidad con otras herramientas: Text-to-Onto es un componente de ambiente integrado para la ingeniería manual y semiautomática de ontologías.

Se puede encontrar información sobre esta herramienta en: *http://ontoserver.aifb.unikarlsruhe.de/texttoonto/*

1.4.2. Terminae

Terminae ha sido desarrollada en el Laboratorio de Informática de Paris-Nord, en la Universidad de Paris-Nord (LIPN). Este sistema integra herramientas lingüísticas y de ingeniería de conocimiento. La herramienta lingüística permite definir formas terminológicas desde el análisis de la ocurrencia de términos en un corpus.

El ontologista analiza los usos del término en el corpus para definir los significados de éste. La herramienta de ingeniería del conocimiento envuelve un editor y un navegador para la ontología. La herramienta ayuda a representar conceptos.

Terminae usa un método para construir conceptos a partir del estudio de las palabras correspondientes a un determinado corpus. Primero, la herramienta establece la lista de vocablos, que requiere la constitución de un corpus relevante en el dominio. Usando una herramienta extractora de términos, un

juego de términos candidatos es propuesto al ontologista, quien selecciona los correctos. Luego los conceptualiza y analiza los usos de cada uno en el corpus para definir todos los significados de éstos. El ontologista además, da la definición en lenguaje natural para cada significado y luego traduce la definición a un lenguaje de implementación *(Gómez, 2003).*

Meta de la herramienta: Construir una ontología.
Técnica de aprendizaje usada por la herramienta: Clusterización conceptual.
Método seguido para el aprendizaje de ontología: Un experto extrae términos desde la lista de términos candidatos y define nociones para los significados de dichos términos.
Intervención usuario/experto en el proceso: Validación.
Interoperabilidad con otras herramientas: Terminae importa la lista de términos extraídos desde la herramienta extractora de términos Lexter[13].

Se puede encontrar información sobre esta herramienta en**:** *http://www-lipn.univ-paris13.fr/~szulman/TERMINAE.html*

1.4.3. OntoLearn

La herramienta OntoLearn está dirigida para extraer términos relevantes de un corpus de texto, relacionarlos a conceptos apropiados en una ontología de propósito general, y detectar relaciones entre los conceptos. Para llevar a cabo estas tareas, son usadas las técnicas de análisis de lenguaje natural y de aprendizaje de máquinas.

OntoLearn extrae terminología de un corpus de texto de dominio, como sitios Web especializados. El sistema entonces filtra los términos usando técnicas de procesamiento de lenguaje natural y estadísticas que lleva a cabo análisis comparativos a través de diferentes dominios. Este análisis identifica la terminología que es usada en el dominio objetivo pero no en otros dominios. Luego, usa la base de conocimiento léxico WordNet[14] para realizar la interpretación semántica de términos. La herramienta entonces relaciona conceptos de acuerdo a relaciones taxonómicas y otras relaciones

[13] ***Lexter*** *es un extractor de términos para el procesamiento de documentos en francés que fue programado en lenguaje C. Fue desarrollado para mantener actualizado un tesauro de términos de la compañía Electricité de France cuyo principal objetivo era la indización automática de documentos (Barrón, 2007).*
[14] ***WordNet****: Es una enorme base de datos léxica del idioma Inglés. Agrupa las palabras en conjuntos de sinónimos llamados 'synsets', proporcionando definiciones cortas y generales, y almacenando las relaciones semánticas entre estos conjuntos de sinónimos (Wikipedia, 2008).*

semánticas, generando un bosque de conceptos de dominio. Para este propósito, WordNet y un método basado en reglas y de aprendizaje inductivo han sido usados para extraer dicha relación. Finalmente OntoLearn integra el bosque de conceptos de dominio con WordNet para crear una vista de la ontología de dominio podada y especializada. La validación del proceso es realizada por un experto *(Gómez, 2003).*

***Meta de la herramienta**:* Enriquecer una ontología de dominio con conceptos y relaciones.
Técnica de aprendizaje usada por la herramienta: Procesamiento de lenguaje natural y aprendizaje de máquinas.
Método seguido por el aprendizaje de ontologías: Interpretación semántica de OntoLearn.
Intervención usuario/experto en el proceso: Validación.

1.4.4. SOAT (Semi-Automatic Domain Ontology Acquisition Tool)

La herramienta SOAT (Wu & Hsu, 2002) permite la adquisición de ontologías de dominio semiautomáticamente de un corpus. El principal objetivo de esta herramienta es extraer relaciones a partir de oraciones parseadas basadas en la aplicación de reglas de frase para identificar palabras clave con fuerte relación semántica, como sinónimos.

El proceso de adquisición está basado en el uso de InfoMap (Hsu et al., 2001), un framework[15] para la representación del conocimiento, que integra conocimiento lingüístico, de sentido común y de dominio. InfoMap ha sido desarrollado para realizar entendimiento de lenguaje natural y capturar las palabras conceptualizadoras, las cuales usualmente son pares formados por sustantivo y adjetivo, o sustantivo y sustantivo. InfoMap tiene dos importantes relaciones entre conceptos: relaciones taxonómicas (categoría y sinónimo) y relaciones no taxonómicas (atributo y evento).

El proceso de adquisición llevado acabo por SOAT incluye coleccionar palabras clave de dominio y encontrar las relaciones entre ellas. Para realizar esta actividad, un juego de reglas ha sido definido para extraer las palabras clave de una oración relacionadas a conceptos en InfoMap con una fuerte

[15] *Un **framework** es una estructura de soporte definida en la cual otro proyecto de software puede ser organizado y desarrollado (Wikipedia, 2008).*

relación semántica entre ellos. La herramienta recibe como entrada un corpus de dominio etiquetado. Una palabra clave, usualmente el nombre del dominio, es seleccionado en el corpus como raíz. Entonces, con esta palabra clave, el proceso tiene por objetivo encontrar una nueva palabra clave relacionada con la anterior mediante la aplicación de reglas de extracción, y adiciona la nueva palabra clave en la ontología de acuerdo con las reglas y la estructura fijada en InfoMap. Esta nueva palabra clave es ahora tomada como raíz para repetir el proceso durante un determinado número de veces o hasta que sea imposible encontrar un nuevo término clave. La intervención del usuario es requerida para verificar los resultados de la adquisición y refinar y actualizar las reglas de extracción.

La restricción de SOAT es que la calidad del corpus debe ser muy alta en el sentido de que las oraciones deben ser lo suficientemente precisas, de manera que incluyan la mayor parte de las relaciones a extraer *(Gómez, 2003).*

Meta de la herramienta: Adquisición de relaciones usando un framework de representación de conocimiento predefinido.
Técnica de aprendizaje de la herramienta: Patrones de expresión, definidos como conjuntos de plantillas lingüísticas.
Método seguido por el aprendizaje de ontología: No especificado, la herramienta sigue su propio método.

Se puede encontrar información sobre esta herramienta en**:** *http://www.iis.sinica.edu.tw/IASL/en/index.htm*

1.5. Herramientas para los Niveles del Análisis Lingüístico

Con objeto de mostrar la complejidad del procesamiento automático o semiautomático del lenguaje natural, se presentan a continuación los diferentes tipos de herramientas necesarias para los distintos niveles de análisis lingüístico.

1.5.1. Analizadores Lexicológicos

En la tarea del tratamiento del lenguaje natural, se reserva para los analizadores lexicológicos un papel muy importante, pues serán los encargados de leer el texto en cuestión para después dividirlo en palabras, que son la unidad mínima con significancia.

En este análisis lingüístico es donde se lleva a cabo el etiquetado de palabras, lo cual trae como resultado la asignación a cada una de las unidades léxicas presentes[16] el conjunto de sus categorías gramaticales posibles. Luego el objetivo de un proceso de etiquetado consiste en la asignación automática de descriptores, etiquetas, o tags, a cada palabra dentro de una oración, donde estas etiquetas corresponden a la categoría gramatical asociada a cada palabra (Valencia, 2005). Es por esto que en la propuesta teórica planteada con esta investigación, se muestran algunas herramientas etiquetadoras para obtener la frase etiquetada con las categorías gramaticales correspondientes de cada palabra.

Las categorías gramaticales han estado presentes en lingüística durante mucho tiempo. Dionysios Thrax (170 aC-90 aC), escritor griego que hizo el notable descubrimiento de que existía la ciencia gramatical, distinguió entre ocho tipos de palabras, usando las siguiente categorías: nombre, verbo, participio, artículo (incluyendo los pronombres relativos), pronombre, preposición, adverbio y conjunción (Valencia, 2005).

Para la obtención de todas las palabras del texto objeto de tratamiento se hace uso de la tokenización, proceso que se podría definir como la segmentación de un texto en palabras-tokens individuales. Los problemas relacionados con la tokenización y el análisis léxico han sido importantes, pero la resolución de la ambigüedad es, para muchos autores, el subproblema más difícil en el análisis.

El problema de la desambiguación consiste en que las palabras tomadas en forma aislada son ambiguas respecto a su categoría. Se puede considerar el siguiente ejemplo *(Valencia, 2005)*:

[16] *Palabras dentro de una oración*

- Hoy **como** pollo. Aquí '**como**' actúa como **verbo**.
- La vida es **como** una caja de bombones. Y en cambio aquí '**como**' se comporta con un **adverbio**.

Afortunadamente la categoría de la mayoría de las palabras no es ambigua dentro de un contexto. Así, con la ayuda de las palabras que no son ambiguas, se podrá inferir la categoría más probable de las ambiguas.

El objetivo de un desambiguador (también llamado *etiquetador morfosintáctico*) es el de asignar a cada palabra la categoría más *apropiada*, dentro de un contexto. Es decir, dada una secuencia de palabras, dotada cada una de un conjunto de etiquetas posibles, el desambiguador debe devolver una secuencia de etiquetas que sea la más *verosímil* dado el contexto. Por supuesto, la calidad del desambiguador depende del grado de precisión (*la granularidad*) del etiquetado del contexto considerado, y de la información de que disponga el desambiguador para considerar *apropiada* una etiqueta o *verosímil* una secuencia de etiquetas. A veces, los desambiguadores no resuelven totalmente el problema de la ambigüedad gramatical y se limitan a eliminar las opciones imposibles o menos probables (Valencia, 2005).

1.5.2. Extractores de Términos

Los extractores de términos son herramientas de explotación de documentos para el establecimiento de un conjunto de términos relevantes, asociados al dominio específico referido en los textos *(Carrera, 2007).*

En la actualidad existen muchas técnicas para la adquisición de términos en corpus lingüísticos (Jacquemin, 2001). La entrada usual a estos sistemas son corpus etiquetados y la salida es una lista con los términos candidatos que aparecen en el corpus.

El proceso de búsqueda es bastante simple y la salida de este proceso es una lista que contiene las expresiones del fragmento de texto actual que aparecen en la base de conocimiento. Esta lista de palabras, la cual se conoce como términos candidatos, deberá ser examinada para determinar si se trata realmente de términos relevantes en el dominio trabajado. Es interesante en este punto

establecer la diferencia entre un término candidato y un término, obteniendo otra definición *(Carrera, 2007)*:

El concepto de término candidato (también colocación) se refiere a un grupo de palabras cuyo significado global se deduce de las unidades que lo componen, por ejemplo: Alcalde de París.

El concepto de término se refiere a una colocación que tiene propiedades sintácticas y representa una realidad en un dominio de conocimiento, por ejemplo: Inteligencia Artificial.

Se trata, pues, de un trabajo de construcción y no de un proceso de descubrimiento de la terminología. Existen, sin embargo, dos características deseables en los términos reconocidos *(Carrera, 2007)*:

- Deben poseer estructura léxica específica en el dominio, y estable en el corpus.

- Deben proporcionar una utilidad demostrable en la aplicación final en la que se va a utilizar el término. En otras palabras, han de ser funcionales para la aplicación que los vaya a explotar.

Es necesario sentar las bases que delimiten al término en un dominio concreto y un área de especialidad.

Esta es la razón por la que la mayoría de los sistemas de adquisición terminológica son semiautomáticos. De este modo al final del proceso de extracción lo que se obtiene es una lista de términos candidatos y no una lista de términos. Será entonces cuando el experto aporte sus conocimientos. Para facilitar la tarea, muchas aplicaciones acompañan los términos con sus contextos de aparición y con información adicional como la frecuencia con que aparecen.

Visto esto, se puede definir la extracción de términos como “la lectura anotada de un corpus textual y la selección de candidatos a términos junto con su contexto” *(Carrera, 2007).*

1.5.3. Analizadores Semánticos

La labor del analizador semántico es decidir cuál de los conceptos es el aplicable en el contexto, es decir, desambiguar aquellas palabras que se puedan proyectar sobre más de un concepto de la ontología *(Moreno, 2000).*

Los analizadores semánticos interpretan el significado de las oraciones, en cierto grado el significado individual de las palabras o de los sintagmas. Algunos de los fenómenos lingüísticos que se pueden tratar conciernen a la identificación de las relaciones del verbo y sus argumentos en una oración, o a la identificación de todos los sinónimos equivalentes de los términos de una oración. La expansión de términos puede realizarse empleando fuentes léxicas como EuroWordNet[17] (lo que disminuye la precisión del resultado); sin embargo, el reto actual consiste en añadir sólo aquellos términos que son pertinentes y que suponen la expansión acertada del significado particular de la palabra en la oración.

Una de las tareas clave de la interpretación semántica consiste en considerar qué combinaciones de significados de palabras individuales son posibles a la hora de crear un significado coherente de la oración, lo que puede reducir el número de posibles significados para cada palabra de una oración determinada.

1.6. Formas de Representación de las Ontologías

El conocimiento inferido se puede representar de diferentes maneras, como figuras, diagramas, etc.; pero la forma más precisa para expresar una ontología, y por ende la más popularizada en el mundo informático es el uso de ficheros XML+ RDF + OWL.

Aquí se muestran algunos ejemplos de representaciones ontológicas.

[17] ***EuroWordNet:*** *(http://www.illc.uva.nl/EuroWordNet/). Es un sistema de redes semánticas para lenguajes europeos (holandés, italiano, español, alemán, francés, checo y estonio) (Wikipedia, 2008).*

1.6.1. Mediante un Fichero XML (Lenguaje OWL)

En este ejemplo tomado de *(Carrera, 2007)* se emplea el concepto "humanidad" para expresar la ontología, también se explica la teoría preliminar del uso de estos lenguajes.

1.6.1.1. Espacio de Nombres

Para poder emplear términos en una ontología, es necesario indicar de qué vocabulario provienen. De este modo, al igual que sucede en muchos documentos XML, una ontología empieza con la declaración del espacio de nombres que se encierra en una etiqueta *rdf*.

```
<rdf:RDF
    xmlns = http://domain.tld/path/humanidad#
    xmlns:owl = http://www.w3.org/2002/07/owl#
    xmlns:rdf = http:// www.w3.org/1999/02/22-rdf-syntax-ns #
    xmlns:rdfs = http:// www.w3.org/2000/01/rdf-schema#
    xmlns :xsd = http:// www.w3.org/2001/XMLSchema#
```

En el primer espacio de nombre se indica el actual. En los cuatro siguientes se indican los espacios de nombre relativos a owl, rdf, rdfs y xsd. Esto es necesario para poder emplear etiquetas propias definidas para esos lenguajes.

1.6.1.2. Cabecera de la Ontología

A continuación de la declaración del espacio de nombres se puede indicar la cabecera que describe el contenido de la ontología actual. Es la etiqueta *owl: Ontology* que permite indicar estas informaciones.

```
<owl:Ontology rdf:about=""
    <rdfs:comment>Ontología que describe la humanidad</rdfs:comment>
…
```

1.6.1.3. Las Clases

Las entidades del "mundo real" que se pueden categorizar en grupos o conjuntos de objetos con características similares, forman las clases de la ontología. Las entidades pueden ser cosas físicas como por ejemplo, los automóviles, o conceptuales como teorías científicas. Éstas constituyen el núcleo de una ontología y describen los conceptos de un dominio concreto.

Un ejemplo de una clase podría ser "Automóvil", que idealmente representaría a todos los automóviles del mundo. De este modo, cada coche sería una instancia o ejemplar de la clase "Automóvil". Una clase puede tener además subclases, que representan conceptos más específicos que el de su superclase. La clase "Automóvil" se podría dividir en las subclases "Turismo", "Todoterreno" y Deportivo".

Una clase puede ser definida de cuatro maneras diferentes:

Indicador de clase:

La declaración se hace nombrando directamente la clase. Este es la única de las cuatro maneras que permite definir un nombre para la clase. En los otros casos las clases definidas se conocen como "anónimas".

```
<owl:Class rdf:ID="Humano"/>
```

Enumeración de los individuos:

Mediante una enumeración de los individuos que pertenecen a una clase se puede definir la clase que los posee. Esto se hace mediante la propiedad *owl: oneOf*.

```
<owl:Class>
    <owl:oneOf rdf:parseType="Collection">
        <owl:Thing rdf:about="María"/>
        <owl:Thing rdf:about="Luis"/>
    </owl:oneOf>
<owl:Class>
```

A través del tipo *Collection* definido en la variable *parseType* de rdf se indica que lo que se presenta es un conjunto de elementos, en este caso de personas.

Restricción de las propiedades:

En este caso se define una clase anónima cuyas instancias deben satisfacer una determinada propiedad. Este tipo de restricciones pueden ser de valor o de cardinalidad. En el primer caso la restricción se limita a un valor de una propiedad del individuo. En el segundo caso la restricción se enfoca al número de valores que puede tomar una determinada propiedad.

```
<owl:Class rdf:ID=”FamiliaNumerosa”>
    <rdfs:subClassOf>
        <owl:Restriction>
            <owl:onProperty rdf:resource=”#tieneNumHijos”/>
            <owl:mincardinality  rdf:datatype=”&xsd;nonNegativeInteger”>
                3</owl:mincardinality>
        </owl:Restriction>
    </rdfs:subClassOf>
</owl:Class>
```

En este caso la clase anónima viene definida entre las etiquetas *subClassOf*. El ejemplo muestra como la clase *FamiliaNumerosa* debe de ser una subclase de una clase anónima que indica que para la propiedad *tieneNumHijos* el valor ha de ser por lo menos 3.

Herencia:

A través de una propiedad de herencia se puede crear una nueva clase. Esto se hace mediante la etiqueta subClassOf.

```
<owl:Class rdf:ID=”Hombre”>
      <rdfs:subClassOf rdf:resource=”#Humano”/>
</owl:Class>
```

1.6.1.4. Las Instancias

Los individuos, instancias o ejemplares, consisten en representaciones de objetos o elementos particulares de una clase. Se denominan indistintamente individuos o instancias de la clase. Hay que señalar que es difícil distinguir entre individuos y clases. Ejemplos de instancias son “España”, “Documento 141203448-5” y “Ford Mustang”.

Un individuo puede hacerse miembro de una clase de diversas maneras:

Mediante una declaración se indica que un determinado individuo pertenece a una clase y se indica además cuales son los valores de propiedad para ese individuo.

```
<Humano rdf:ID=”María”>
    <tienePorMadre rdf:resource=”#Rosa”/>
    <tienePorPadre rdf:resource=”#Miguel”/>
</Humano>
```

Mediante una declaración anónima para ello se procede como en el caso anterior pero obviando el nombre del individuo.

Estableciendo una relación entre el individuo y otro ya existente. Esto se consigue utilizando las propiedades, *owl:sameAs owl:differentFrom* y *owl:allDifferent* .

```
rdf:about=”Jose”>
    <owl:sameAs rdf:resource=”Pepe”/>
</rdf:Description>
```

1.6.1.5. Las Propiedades

Las entidades que pertenecen a una clase poseen atributos determinados, por ejemplo, tienen un nombre, un color o un peso. Por tanto, las propiedades consisten en pares de atributo/valor y sirven para describir de forma conveniente las características relevantes de las entidades que forman las clases. Algunos ejemplos son “Población”, “ISBN” y “Precio”.

Las propiedades permiten expresar hechos sobre las clases y sus instancias.

OWL distingue entre dos tipos de propiedades:

- **Propiedades de objeto** que permiten unir instancias de clase. Este tipo de propiedad es una instancia de la clase *owl: ObjectProperty*.
- **Propiedades de tipo de dato**, que permiten enlazar individuos con valores. Este tipo de propiedad es una instancia de la clase *owl: DataTypeProperty*.

```
<owl:ObjectProperty rdf:ID=”tienePorPadre”/>
<owl:DataTypeProperty rdf:ID=” tieneNumHijos”/>
```

Es posible asignar otro tipo de características a las propiedades, entre las que se encuentran la transitividad, la simetría, etc.
Especificar una propiedad no es más que restringir la relación que simboliza. Así, deben verse las propiedades como una función que hace corresponder a un individuo otro individuo o bien un valor. Por lo tanto, se pueden definir el dominio y la imagen de esta propiedad. Además se puede definir una propiedad como una especialización de una existente. Como se muestra en el siguiente ejemplo:

```
<owl:ObjectProperty rdf:ID=”vive”>
    <rdfs:domain rdf:resource=”#Humano”/>
    <rdfs:range rdf:resource=”#País”/>
</owl:ObjectProperty>
```

tanto Humano como País son clases que han sido definidas en la ontología.

1.6.2. Mediante un Gráfico

Otra forma de representar las ontologías es mediante un gráfico, como se muestra en la Figura 2, donde se representa un ejemplo de clases y subclases de una ontología de periféricos de ordenador:

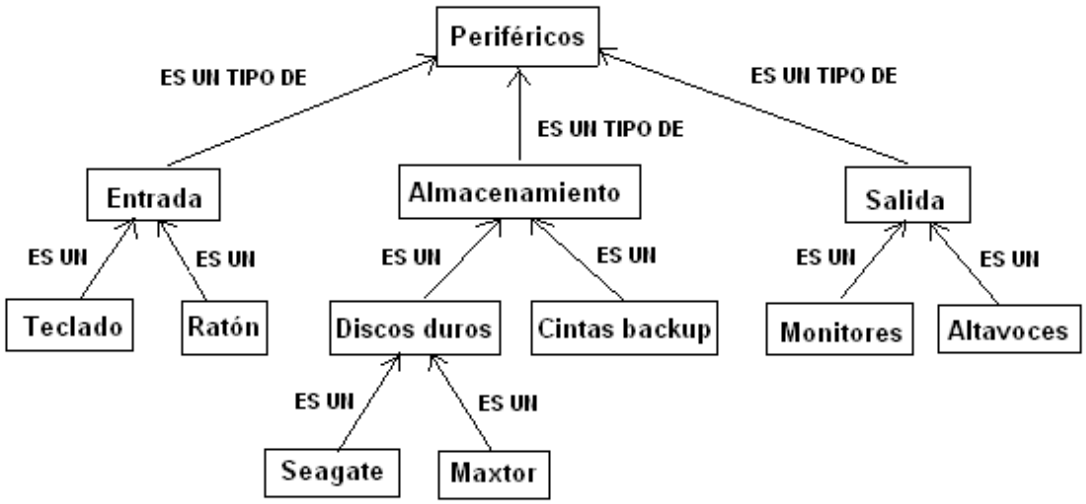

Figura 2: Clases y subclases de una ontología sobre periféricos de ordenador (Codina, 2007).

A continuación, en la Figura 3, se muestra parte de la ontología anterior representada mediante el lenguaje OWL.

```
<?xml version="1.0" ?>
<rdf: RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
  ...
<owl:Class rdf:ID="Perifericos">
 <rdf:comment>
Los periféricos de ordenador están conectados a la CPU pero no forman parte de ella.
 </rdf:comment>
</owl:Class>

<owl:Class rdf:ID="Entrada">
 <rdf:comment>
Los periféricos de entrada son una subclase de periféricos de ordenador.
 </rdf:comment>

 <rdfs:subClassOf rdf:resource="#Perifericos" />
</owl:Class>

<owl:Class rdf:ID="Teclados">
  <rdf:comment>
Los teclados son una subclase de los periféricos de entrada.
 </rdf:comment>
  <rdf:subClassOf rdf:resource="#Entrada" />

  <rdf:subClassOf rdf:resource="#Perifericos" />
</owl:Class>
  ...
</rdf:RDF>
```

Figura 3: Ontología representada mediante el lenguaje OWL (Codina, 2007).

1.7. El e-Learning

El concepto de e-Learning es comprendido fácilmente por la mayoría de las personas que se relacionan con el mundo de la informática. E-Learning implica una nueva forma de aprender. No quiere decir que es totalmente diferente a la enseñanza tradicional en la escuela, en la universidad o durante nuestra vida laboral, sino que integra nuevas fuentes y formas de adquirir conocimiento. Actualmente existen muchas variantes para la concepción del aprendizaje electrónico, a continuación se muestran algunas de las definiciones más comunes:

Enseñanza a distancia caracterizada por una separación física entre profesorado y alumnado (sin excluir encuentros físicos puntuales), entre los que predomina una comunicación de doble vía asíncrona donde se usa preferentemente Internet como medio de comunicación y de distribución del conocimiento, de tal manera que el alumno es el centro de una formación independiente y flexible, al

tener que gestionar su propio aprendizaje, generalmente con ayuda de tutores externos *(Wikipedia, 2008).*

Técnicamente, el e-Learning es la entrega de material educativo vía cualquier medio electrónico, incluyendo el Internet, Intranets, Extranets, audio, vídeo, red satelital, televisión interactiva, CD y DVD, entre otros medios *(Mendoza, 2003).*

Algunos autores han acotado el alcance del e-Learning, como (Red TTnet, 2005) que lo define como el "conjunto de tecnologías, aplicaciones y servicios orientados a facilitar la enseñanza y el aprendizaje a través de Internet/Intranet, que facilitan el acceso a la información y la comunicación con otros participantes."

La educación electrónica es la capacitación y adiestramiento de personas usando las comodidades digitales y en general Internet, llegando a ofrecer sofisticadas facilidades como flujo de audio y vídeo, presentaciones, vínculos a información relativa al tema que se trate. Y en fin, cualquier forma de capacitar y enseñar usando las TIC[18].

Las herramientas que componen esta estrategia de educación son, por un lado, diferentes utilidades para la presentación de los contenidos (textos, animaciones, gráficos, vídeos) y por otro, herramientas de comunicación entre alumnos o entre alumnos y tutores de los cursos (correo electrónico, chat, foros). Pero, más allá de las herramientas ocupadas, el e-Learning, como todo proceso educativo, requiere de un diseño instructivo sólido y que tome en cuenta, además de las consideraciones pedagógicas, las ventajas y limitaciones de Internet y el comportamiento de los usuarios de la misma.

1.7.1. Objetos de Aprendizaje

En la misma medida que la información en la Web se incrementa, también aumentan los recursos que pueden utilizarse en el sector educativo. El término "objeto de aprendizaje" ha surgido con la finalidad de compartir recursos y reutilizarlos en el dominio del e-Learning.

A continuación se muestran algunas definiciones sobre los OAs:

[18] ***TIC:*** *Tecnologías de la Información y la Comunicación*

Un objeto de aprendizaje es "Cualquier recurso con una intención formativa, compuesto de uno o varios elementos digitales, descrito con metadatos, que pueda ser utilizado y reutilizado dentro de un entorno e-Learning" (López, 2005).

Los objetos de aprendizaje son materiales educativos que tienen un objetivo determinado, un contenido y un grupo de actividades de aprendizaje. Se generan bajo estándares que permiten que se puedan empaquetar y trasladar a otras aplicaciones desde las cuales se almacenan (Repositorios de Objetos de Aprendizaje) y se comparten con otros usuarios a través de sus características de empaquetado y metadatos *(Aprendercontics).*

Los objetos de aprendizaje son tan variados como gráficos, fotos, imágenes, videos, textos, sonidos, capítulos de cursos o incluso cursos completos, por lo que puede considerarse difícil organizar todos estos archivos de manera eficiente y devolver resultados útiles ante la búsqueda de información en los repositorios que almacenan dichos recursos.

Algunas características generales que presentan los OA son las siguientes *(García, 2005)*:

- ***Reusabilidad:*** Capacidad para ser usado en contextos y propósitos educativos diferentes y para adaptarse y combinarse dentro de nuevas secuencias formativas.

- ***Educatividad:*** Capacidad para generar aprendizaje.

- ***Interoperabilidad:*** Capacidad para poder integrarse en estructuras y sistemas (plataformas) diferentes.

- ***Accesibilidad:*** Facilidad para ser identificados, buscados y encontrados gracias al correspondiente etiquetado a través de diversos descriptores (metadatos), que permitirían la catalogación y almacenamiento en los repositorios.

- ***Durabilidad:*** Vigencia de la información de los objetos, sin necesidad de nuevos diseños.

- ***Independencia y autonomía*** de los objetos con respecto a los sistemas desde los que fueron creados y con sentido propio.

- ***Generatividad:*** Capacidad para construir contenidos, objetos nuevos derivados de él. Capacidad para ser actualizados o modificados, aumentando sus potencialidades a través de la colaboración.

- ***Flexibilidad, versatilidad y funcionalidad:*** Elasticidad para combinarse en diversas propuestas de áreas del saber diferentes.

1.7.2. Los Metadatos

Los metadatos surgieron de la necesidad de describir, identificar y caracterizar objetos de información (documentos impresos, sonoros, etc.) a través de campos que identifican cada dato con el propósito de facilitar la gestión de los documentos. Ejemplo: título, autor, materia, resumen, editor, entre otros.

Actualmente la utilización de metadatos ha logrado difundirse con rapidez y en estrecha relación con los nuevos objetos de información, como son los recursos electrónicos generados y disponibles en Internet. Por ello la definición de metadatos cobra un nuevo sentido:

"Información que describe datos que incluyen el contenido, la forma y las características técnicas y editoriales de la información electrónica, los cuales son generados, consultados, manipulados y distribuidos en la red" *(Sánchez, 2002).*

Otra definición de metadatos sería: "Información estructurada que describe, explica, localiza, o maneja información especializada acerca de un recurso de información que tenga algún sentido para las máquinas o los humanos" *(DPCBD, 2006).*

Una de las características más importantes de los metadatos es su capacidad de relación o de establecer enlaces, de esta forma se han hecho imprescindibles en la recuperación global de la información en Internet, puesto que se trata de indizar y clasificar inconmensurables cantidades de información de diversos tipos. Se tratará de integrar de forma heterogénea fuentes de información muy diversas, así como integrar diferentes formatos de bases de datos. Dicho de otra manera, se emplean metadatos para organizar el contenido de la información en Internet. De esta forma en la definición de

metadatos se puede incluir y de forma más específica, que se refiere a la información accesible por Internet, por lo que los metadatos tienen el objetivo primordial de que los documentos introducidos en la red incluyan todos los datos necesarios para su posterior búsqueda, localización y recuperación. Ya que introducir o publicar dentro de Internet es una tarea sencilla, sin embargo la localización, control y uso de la información es una tarea más compleja. Por tanto, será una tarea primordial establecer las normas y elementos que ha de contener cualquier descripción y catalogación de recursos en Internet *(Adrián, 2007).*

Funciones de los metadatos *(Sánchez, 2002)*:

- Proporcionar una descripción de un objeto de información junto con otra información necesaria para su manejo y preservación.
- Proporcionar los puntos de acceso a esa descripción.
- Codificar esa descripción.

Propósito de los metadatos *(Sánchez, 2002)*:

- Para ayudar en la determinación de la relevancia de una información relacionada con una consulta concreta.
- Para analizar las características de un conjunto de información
- Para enviar información de un sistema a otro diferente.

Los metadatos son empleados junto a los OA para garantizar su reutilización pues posibilitan la comunicación declarando cómo están relacionados los datos y convirtiéndolos de un formato a otro, posibilidad que es empleada en la representación del conocimiento y en las consultas a buscadores donde se obtienen resultados más precisos.

Los metadatos aplicados a objetos de aprendizaje permiten *(Rodríguez, 2007)*:

- Crear descripciones bien estructuradas de recursos de aprendizaje, que faciliten el descubrimiento, localización, evaluación y adquisición de recursos de aprendizaje por parte de estudiantes, profesores o procesos de software automáticos.

- Compartir descripciones de recursos entre sistemas de búsqueda de recursos, que propicien una reducción en los costes de los servicios proveedores basados en descripciones de recursos de alta calidad.

- Adaptación de las descripciones de recursos a las necesidades concretas de una comunidad, que puede incluir la elección de vocabularios controlados y adecuados para la clasificación, reducción o ampliación del número de elementos de descripción.

1.8. Los Repositorios

Un repositorio es una colección de recursos digitales, accesibles a través de una red. Pueden incluir los recursos, los metadatos que describan dichos recursos o ambos (Bianciotto, 2008).

El origen de la palabra española repositorio deriva del latín ¨repositorium¨, que significa armario, alacena, y se aplicó al léxico específico de la informática para designar los depósitos de información digital, los cuales pueden ser de acceso público, o pueden estar protegidos y necesitar de una autentificación previa. La propia web puede considerarse como un gran repositorio, siempre que se le apliquen las estrategias de búsqueda, procesamiento, selección y catalogación a través de esquemas de metadatos

Los repositorios más conocidos son los de carácter académico y los institucionales, estos se clasifican según el tipo de información que manejan, por ejemplo, existen repositorios de actualizaciones de sistemas operativos (Debian, Gentoo, etc.), de bases de datos, de programas, de objetos de aprendizaje, entre otros, los cuales permiten a los usuarios la obtención de recursos a través de la búsqueda de información mediante los navegadores web.

1.8.1. Repositorios de Objetos de Aprendizaje

Actualmente, la producción de OA y materiales educativos en formato digital, así como su almacenamiento en repositorios cada vez más extensos, es una práctica generalizada en todas las instituciones dedicadas a la formación.

Los repositorios de objetos de aprendizaje (ROA) son comparables con bibliotecas digitales combinado con un buscador de elementos en ellos. Estos repositorios permiten almacenar, buscar, recuperar, consultar y descargar objetos de aprendizaje de todas las áreas del conocimiento provenientes de disímiles fuentes, proporcionando algún tipo de interfaz de búsqueda de los mismos, bien para interacción con humanos o con otros sistemas software.

Estos repositorios agrupan y almacenan los OA de dos formas: la primera, los recursos educativos y sus correspondientes metadatos dentro de un mismo sistema e incluso dentro de un mismo servidor, y la segunda, sólo se encuentran los metadatos, en este caso el repositorio contiene sólo los descriptores, y se accede al objeto a través de una referencia a su ubicación física que puede ser en otro sistema o repositorio de objetos. En ambos casos el contenido puede ser accedido por otros sistemas o usuarios mediante interfaces de búsquedas *(López, 2006).*

En la actualidad existe una tendencia a crear enormes redes de repositorios locales (Surós & Pernía, 2006):

- Merlot (Multimedia Educational Resource for Learning and Online Teaching)
- Careo (Campus Alberta Repository of Educational Objects)
- MarcoPolo
- Alejandría
- Morea (Múltiples Objetos Reutilizables para la Enseñanza y el Aprendizaje)
- Color (Colección de Objetos Reusables)
- Aproa (Aprendiendo con Objetos de Aprendizaje)
- ROA (Repositorio de Objetos de Aprendizaje)

En el Polo Productivo de Teleformación perteneciente a la UCI, donde surge el presente trabajo, existe un equipo de investigación y desarrollo encargado de perfeccionar y mejorar la herramienta ROA (Surós & Pernía, 2006), esta herramienta es una aplicación Web que permite almacenar paquetes SCORM[19]. Estos paquetes representan el conjunto de especificaciones que permiten desarrollar, empaquetar y entregar materiales educativos de alta calidad en el lugar y momento necesarios, con la finalidad de que sean interoperables.

[19] ***SCORM:*** *Modelo Referenciado de Objetos de Contenido Compartible*

1.8.2. Repositorios Semánticos de Objetos de Aprendizaje

Para garantizar la reutilización de los objetos y recursos educativos en nuevos contextos de aprendizaje se precisa desarrollar sistemas, cada vez más sofisticados y basados en el uso de la semántica, con objeto de facilitar la gestión y categorización de las extensas colecciones de OA contenidos en la multiplicidad de repositorios que existen, tomando como punto de partida las descripciones cualitativas estimadas por los usuarios, para que respondan con mayor fiabilidad a lo que en realidad quieren referirse *(Cernea, 2007).*

Los repositorios semánticos son capaces de analizar la semántica (significado) de cada uno de los objetos de aprendizaje que los componen y organizar la gestión de objetos de acuerdo a ese análisis. Estos repositorios permiten almacenar cualquier información referente al OA y proporcionan una serie de funcionalidades adaptadas a la conceptualización particular del mismo. De esta manera se le puede brindar al usuario una forma más efectiva de navegación y exploración de toda la información disponible sobre el tópico de su interés.

Con respecto a la búsqueda de información, la búsqueda por palabras clave cuando no se tiene en cuenta su semántica, puede resultar en la recuperación de muchos documentos no deseados. Las ontologías representan la tecnología adecuada para facilitar la búsqueda semántica y proveer "entendimiento" automático de la información, facilitando la explotación del conocimiento contenido en documentos y repositorios de una forma más eficiente *(Martín & Olsina).*

Actualmente en la UCI, se realiza la investigación y desarrollo de un repositorio semántico de OA, para lo cual se destinó un equipo de trabajo compuesto por 5 estudiantes. El presente trabajo se propone aportar una teoría para contribuir con la realización del mismo, haciendo más efectiva la creación de las ontologías de forma automática y semiautomática.

1.9. Conclusiones del Capítulo

En este capítulo se abordaron los fundamentos teóricos relacionados con la generación automática y semiautomática de ontologías de dominio, basados en el procesamiento de lenguaje natural. Se

mostraron las características de un grupo de sistemas y lenguajes especializados en el desarrollo de esta tarea, y se describieron de forma general las herramientas para los niveles del análisis lingüístico a tener en cuenta en la propuesta teórica desarrollada en el capítulo 2. También se expusieron diferentes puntos de vistas sobre la representación de las ontologías, por ejemplo, a través de gráficos y de lenguajes ontológicos. Se explicaron además, un grupo de conceptos relacionados con el dominio en que se enmarca este trabajo, como los repositorios de objetos de aprendizaje, ya que en un futuro, la propuesta que se plantea será aplicada a los mismos, de forma tal que se conviertan en repositorios semánticos.

Capítulo 2: Propuesta Metodológica

2.1. Introducción

En este capítulo se presenta una secuencia de fases para generar automática y semiautomáticamente ontologías de dominio, para de esta forma sentar las bases de la futura creación de un repositorio semántico comenzando por el tratamiento de documentos de texto. También se muestran posibles herramientas a utilizar en cada una de las fases que componen la propuesta teórica, centrándose la misma en el procesamiento de lenguaje natural, y se emiten valoraciones sobre esas herramientas.

2.2. Puntos de Partida

Como se ha visto en el Capítulo 1, con el interés de compartir recursos y para su reutilización en el ámbito educativo ha surgido el concepto de "objeto de aprendizaje", con la finalidad de maximizar el número de situaciones educativas en las que el recurso pueda ser utilizado.

Al incorporar un OA a un repositorio de objetos de aprendizaje se debe tener en cuenta la correcta descripción del mismo, a través de sus correspondientes metadatos, ya que son componentes importantes para su reutilización; pero además, gracias a los metadatos se pueden analizar los diferentes tipos de recursos educativos que se encuentran en el repositorio, ya sean videos, presentaciones, gráficos, cursos, documentos de textos, etc. Esto es imprescindible a la hora de construir una ontología según el tipo de recurso educativo subido al repositorio, ya que la propuesta que se identifica solo tratará los OA calificados como documentos de texto, puesto que existe gran cantidad de bibliografía referente al PLN[20]. Por tanto, se comenzará haciendo una propuesta para la generación automática y semiautomática de ontologías de dominio a partir de texto, la cual fue tomada de (Valencia, 2005), ya que permite simplificar los procesos de adquisición de conocimiento y no sólo se centra en la obtención de jerarquías de conceptos, sino que también tiene en cuenta un amplio conjunto de relaciones semánticas entre entidades de conocimiento.

[20] ***PLN:*** *Procesamiento del Lenguaje Natural*

Esta propuesta se compone de una secuencia de fases que trata de descubrir términos que representen información válida mediante la búsqueda en una base de conocimiento, que relacione reglas y conceptos con expresiones lingüísticas que puedan representar esos conceptos. Además, permite identificar relaciones entre los conceptos, así como inferir nuevas entidades de conocimiento. La idea principal en la que se basa esta propuesta es el hecho que las relaciones entre términos, están asociadas normalmente a expresiones verbales. Es por eso que la propuesta que se presenta permite almacenar las expresiones verbales que representan relaciones entre conceptos, con la finalidad de ser capaz de identificar automáticamente este conocimiento cuando reaparezca (Valencia, 2005).

El mecanismo automático y semiautomático de generación de ontologías identificado por (Valencia, 2005), propuesta que se recomienda en este trabajo, tiene en cuenta los siguientes aspectos:

- Las ontologías pueden utilizarse para representar conocimiento. Como el texto completo puede ser muy largo sería conveniente dividirlo en fragmentos más pequeños para facilitar su procesamiento en frases.

- Los expertos en un dominio pueden extraer conocimiento a partir de un texto de ese dominio.

- Los textos contienen conocimiento. Pueden aparecer entidades de conocimiento a lo largo del texto de una manera explícita, aunque a veces el conocimiento puede estar expresado de manera implícita. (El mecanismo automático y semiautomático que se presenta para generar ontologías, tratará de encontrar sólo el conocimiento explícito en el texto).

2.3. Fases de la Propuesta

La propuesta que se presenta se divide principalmente en tres fases secuenciales (Figura 4): Fase de Etiquetado, Fase de Búsqueda de Conceptos y Fase de Inferencia.

Como se observa en la Figura 4, este mecanismo está supervisado por un experto o un ingeniero de conocimiento para la construcción semiautomática de ontologías. El futuro sistema a construir que utilice esta propuesta irá mostrando al usuario el conocimiento inferido de la frase actual y el usuario

aceptará o modificará ese conocimiento, generando de manera semisupervisada una ontología del dominio.

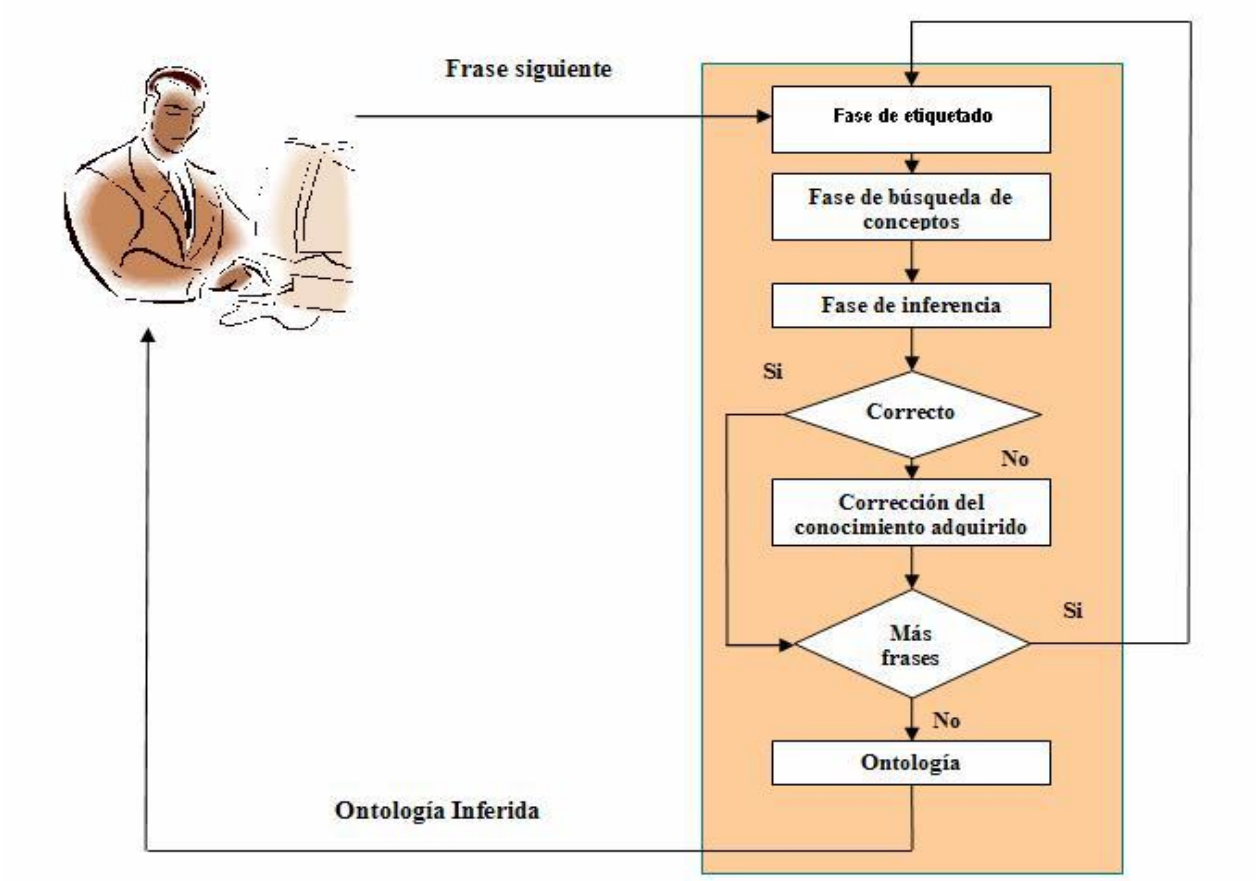

Figura 4: El proceso de adquisición de conocimiento (Valencia, 2005).

A continuación se explican cada una de las fases que componen la propuesta teórica.

2.3.1. Fase de Etiquetado (Tagging)

Los analizadores lexicológicos (ver acápite 1.5.1) son programas que reciben como entrada un texto y producen una salida compuesta de tokens, proporcionando una categoría léxica a las palabras que

componen el texto. Su funcionamiento se puede basar en conocimiento lingüístico o en la aplicación de probabilidades.

Seguidamente se muestran varias herramientas etiquetadoras que se pueden utilizar en esta fase según las características del entorno donde se aplique la propuesta, como parte del proceso de construcción automática o semiautomática de ontologías de dominio.

2.3.1.1. FreeLing

El software FreeLing está diseñado para ser usado como una biblioteca externa desde cualquier aplicación que requiera este tipo de servicios. Está programado en su totalidad usando el lenguaje C++. Esta herramienta fue creada en la Universidad Politécnica de Cataluña, y ha sido mejorada y ampliada gracias a diversas contribuciones hechas.

Algunos servicios que ofrece la herramienta FreeLing *(FreeLing Home Page)*:

- Tokenización de texto.
- División de cadenas.
- Análisis morfológico.
- Tratamiento de sufijos, retokenización de pronombres.
- División de contracciones.
- Predicción probabilística de categorías desconocidas de palabras.
- Detección de entidades nombradas.
- Reconocimiento flexible de palabras múltiples.
- Reconocimiento de fechas, números, magnitudes físicas como velocidad, peso, temperatura, densidad, etc.
- Etiquetación de texto hablado, etc.

La mayoría de estos servicios se prestan actualmente para los siguientes idiomas: español, catalán, italiano e inglés.

Se pueden descargar varias versiones de esta herramienta en la siguiente dirección Web: *https://lafarga.cpl.upc.edu/frs/?group_id=32*.

Este software está orientado para ser usado en sistemas Linux/Unix, licenciado bajo los términos de la licencia GPL, aunque se ha reportado éxito en compilaciones para MacOS. Existen versiones hechas por terceros destinadas para Windows, éstas se suministran por conveniencia, pero sin soporte ni garantía *(FreeLing Home Page).*

Debido a esto se explican sólo los aspectos preliminares de la instalación para Linux, para lo cual es necesario un grupo de requerimientos:

- Un Linux típico con las herramientas de desarrollo clásicas, bash, make, y un compilador de C++ con soporte básico STL.
- 120 MB de disco duro.
- Librerías externas:

    - libpcre (versión 4.3 o mayor) Expresiones regulares de Perl. Disponible en: *http://www.pcre.org*.

    - libdb (versión 4.1.25 o mayor) Incluido en todas las distribuciones usuales de Linux. Debe asegurarse que está instalada con soporte para C++. Disponible en *http://www.sleepycat.com.*

    - libcfg+ (versión 0.6.1 o mayor) Administración y ficheros de configuración con las opciones de la línea de comandos. Disponible en *http://www.platon.sk/projects/libcfg+.*

    - Omlet and Fries (libomet versión 0.97 o posterior) Librerías de utilidades de machine learning.

Para evitar problemas adicionales se recomienda instalar las librerías mencionadas (si no están por defecto en el sistema) en los directorios por defecto, que son /usr/lib o /usr/local/lib para las librerías y /usr/include o /usr/local/include para las cabeceras.

Algunos detalles adicionales de la instalación aquí: *http://garraf.epsevg.upc.es/freeling/index.php?option=com_content&task=view&id=21&Itemid=52*

Posterior a su instalación será necesaria la familiarización del usuario con el sistema, para lo que se presenta un manual para su uso en completo detalle en: *http://garraf.epsevg.upc.es/freeling/doc/userman/html/.*

Ejemplo de una versión online de la herramienta FreeLing en: *http://garraf.epsevg.upc.es/freeling/demo.php*

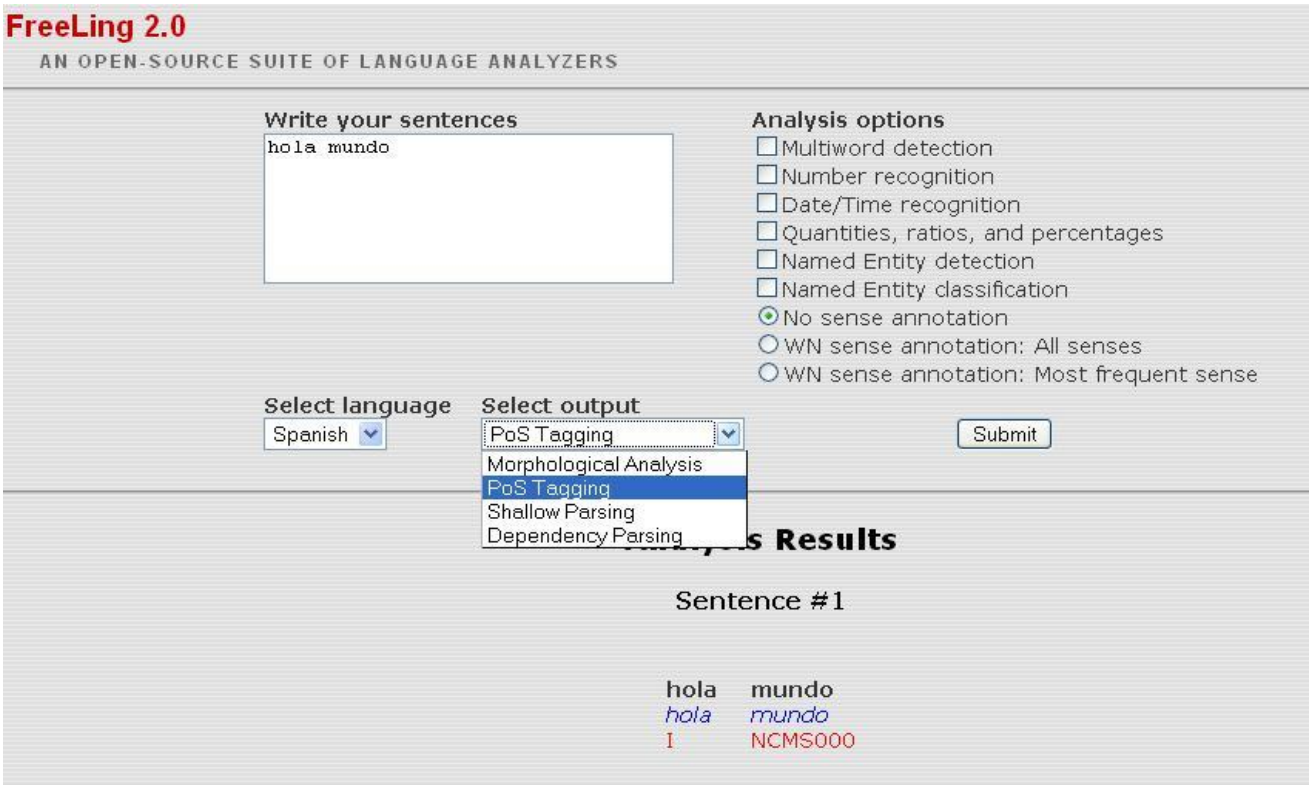

Figura 5: Versión 2.0 de la herramienta FreeLing online (FreeLing Home Page).

2.3.1.2. TreeTagger

Dentro de los etiquetadores estadísticos[21] se encuentra el TreeTagger *(TreeTagger)*, con diccionarios para el francés y el español entre otros. Incluye además un módulo de segmentación que sitúa cada palabra, con su correspondiente análisis, en una línea distinta.

Este programa fue desarrollado por Helmut Schmid en la Universidad de Stuttgart (Instituto de Procesamiento de Lenguaje Natural) en 1994, y está basado en árboles de decisión[22] (Figura 6). El árbol de decisión empleado por esta herramienta es binario. El recorrido del mismo en busca de la etiqueta más adecuada se realiza tomando en cada nodo el camino más probable hasta llegar a un nodo hoja. En cada nodo hoja existe información probabilística que ayuda a la toma de decisión.

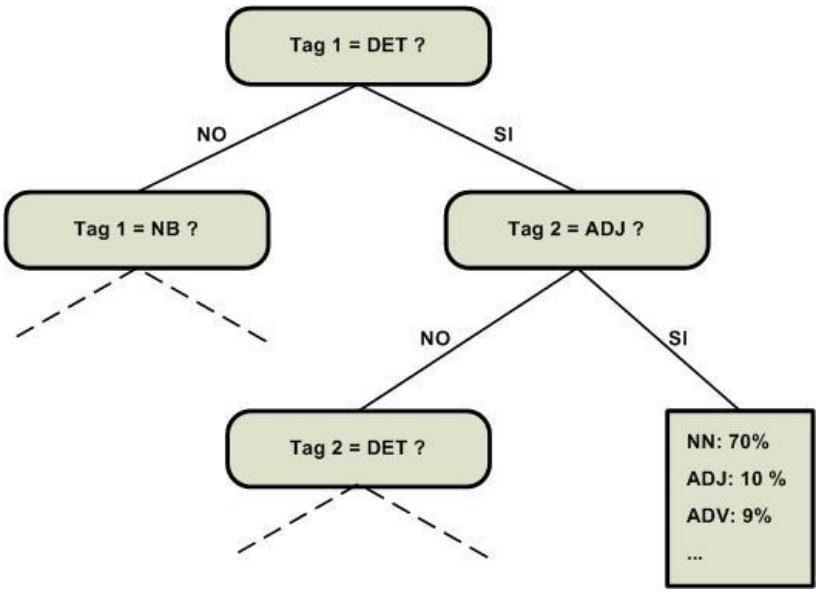

Figura 6: Árbol de decisión en TreeTagger (Carrera, 2007).

[21] ***Etiquetadores estadísticos:*** *etiquetadores que obtienen los modelos del lenguaje y generalizaciones en que basan su actuación automática a partir de la evidencia empírica obtenida de corpus lingüísticos voluminosos (Valencia, 2005).*

[22] *Un **árbol de decisión** es un modelo de predicción utilizado en el ámbito de la inteligencia artificial, dada una base de datos se construyen estos diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que suceden de forma sucesiva, para la resolución de un problema (Wikipedia, 2008).*

Principales características del TreeTagger *(TreeTagger)*:

1. Fácilmente adaptable a idiomas si están disponibles un lexicón y un corpus de entrenamiento manualmente etiquetado.

2. Se ha probado con resultados satisfactorios en alemán, inglés, francés, italiano, español, búlgaro, ruso, griego, portugués, chino.

3. Código ejecutable disponible para Linux. MAC (OS), Windows, y SPARC Workstations.

4. Se pueden especificar incluso nuevos idiomas, distintos a los convencionales y definidos por el usuario.

Durante los años 2007-2008 se ha venido desarrollando la versión más reciente de la herramienta en cuestión. A continuación se muestran dos figuras correspondientes a las interfaces de este software para el sistema operativo Windows.

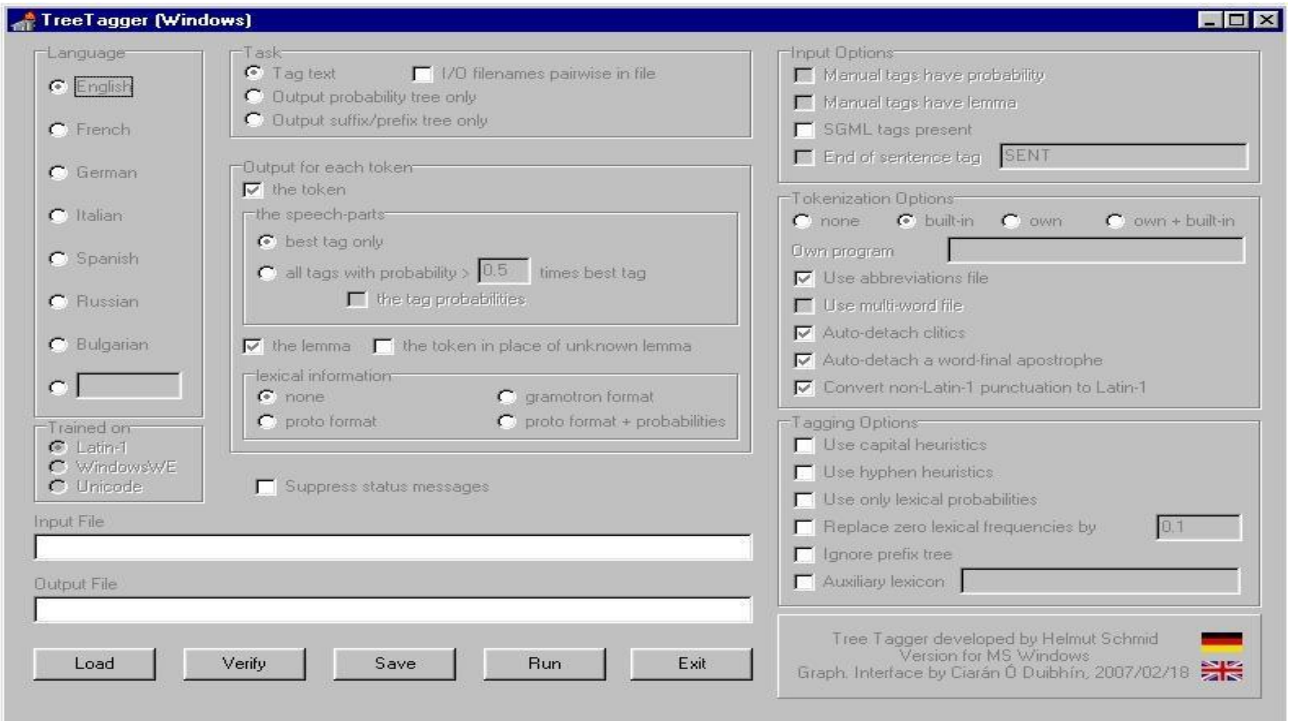

Figura 7: Interfaz de la herramienta TreeTagger para Windows (Ciarán Ó Duibhín, 2008).

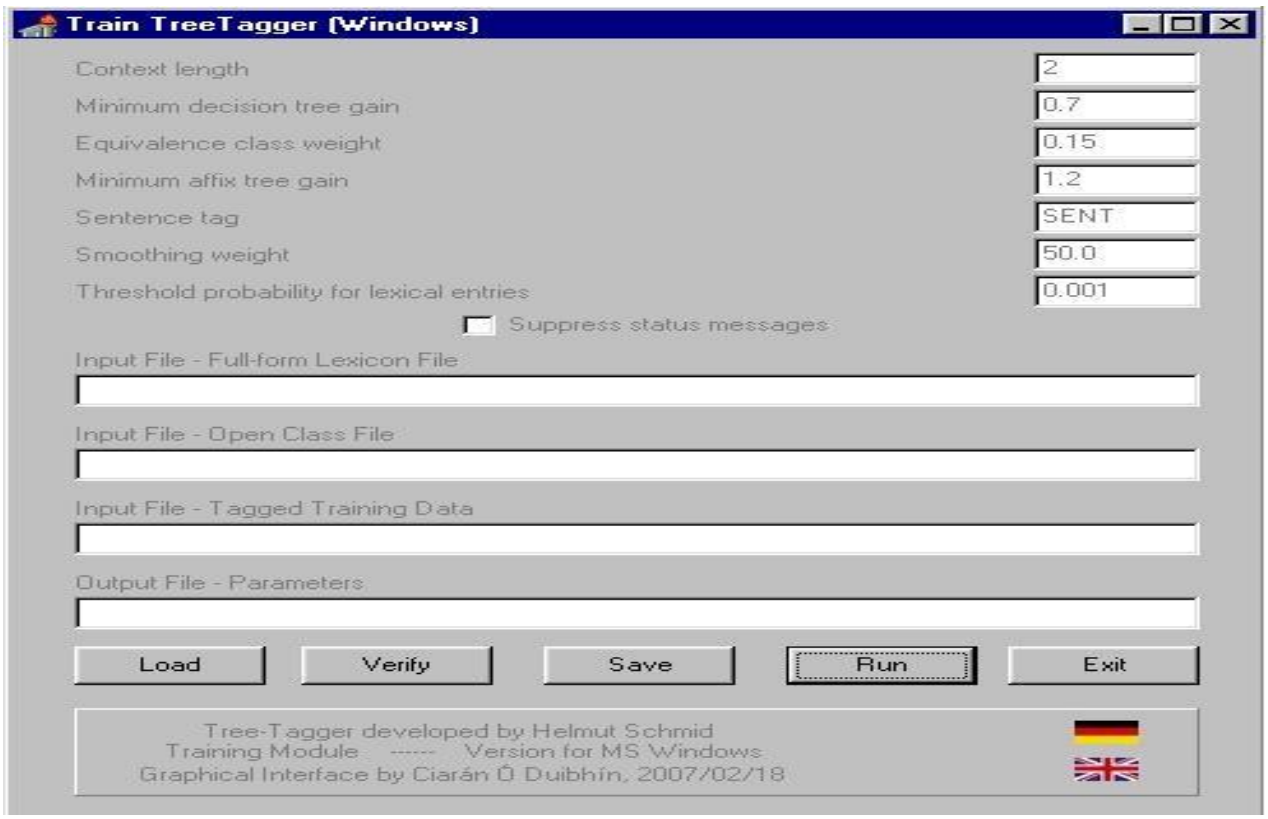

Figura 8: Interfaz de entrenamiento de la herramienta TreeTagger para Windows (Ciarán Ó Duibhín, 2008).

Pasos necesarios para la instalación del TreeTagger *(TreeTagger)*:

1. Descargar el paquete TreeTagger en un determinado directorio:

    - PC-Linux: *ftp://ftp.ims.uni-stuttgart.de/pub/corpora/tree-tagger-linux-3.1.tar.gz*
    - Sparc-Solaris: *ftp://ftp.ims.uni-stuttgart.de/pub/corpora/tree-tagger-3.1.tar.gz*
    - Mac OS-X: *ftp://ftp.ims.uni-stuttgart.de/pub/corpora/tree-tagger-MacOSX-3.1.tar.gz*
    - Windows: *ftp://ftp.ims.uni-stuttgart.de/pub/corpora/tree-tagger-windows-3.1.zip*

2. Descargar los tagging scripts en el mismo directorio.

    ftp://ftp.ims.uni-stuttgart.de/pub/corpora/tagger-scripts.tar.gz

3. Descargar el scritp de instalación install-tagger.sh.

    ftp://ftp.ims.uni-stuttgart.de/pub/corpora/install-tagger.sh

4. Descargar los ficheros de parámetros:

    - PC-Linux : *http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html#Linux*
    - Sparc-Solaris: *http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html#Solaris*
    - Mac: *http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html#Solaris*

5. Abrir una ventana terminal y ejecutar el script de instalación en el directorio donde han sido descargado los ficheros.

    sh install-tagger.sh

6. Hacer una prueba:

    echo 'Hello world!' | *cmd/tree-tagger-english*
    or
    echo 'Das ist ein Test.' | *cmd/tagger-chunker-german*

7. Si existen dificultades con la instalación, dirigirse a la siguiente dirección Web:
   http://www.smo.uhi.ac.uk/~oduibhin/oideasra/interfaces/winttinterface.htm

Además, existen incluso versiones online de este etiquetador. En la Figura 9 se muestra la interfaz de una edición online de esta herramienta que puede ser encontrada en *http://www.cele.nottingham.ac.uk/~ccztk/treetagger.php.*

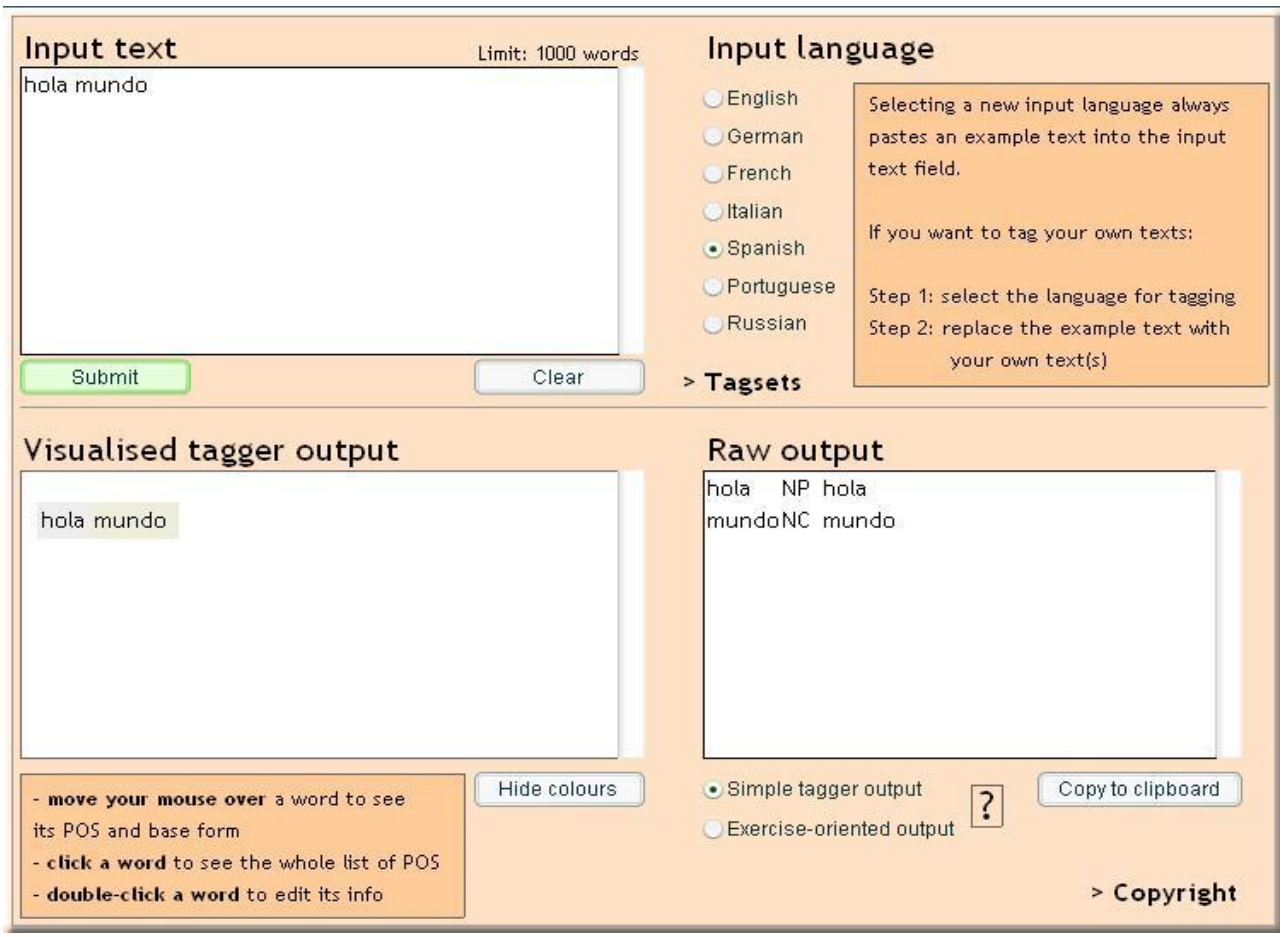

Figura 9: Interfaz de una versión online de la herramienta TreeTagger

A continuación se muestra la etiquetación de una frase en idioma español. En este caso el token NC hace referencia a un sustantivo común, por ejemplo mesa, mesas, libro, libros, etc.

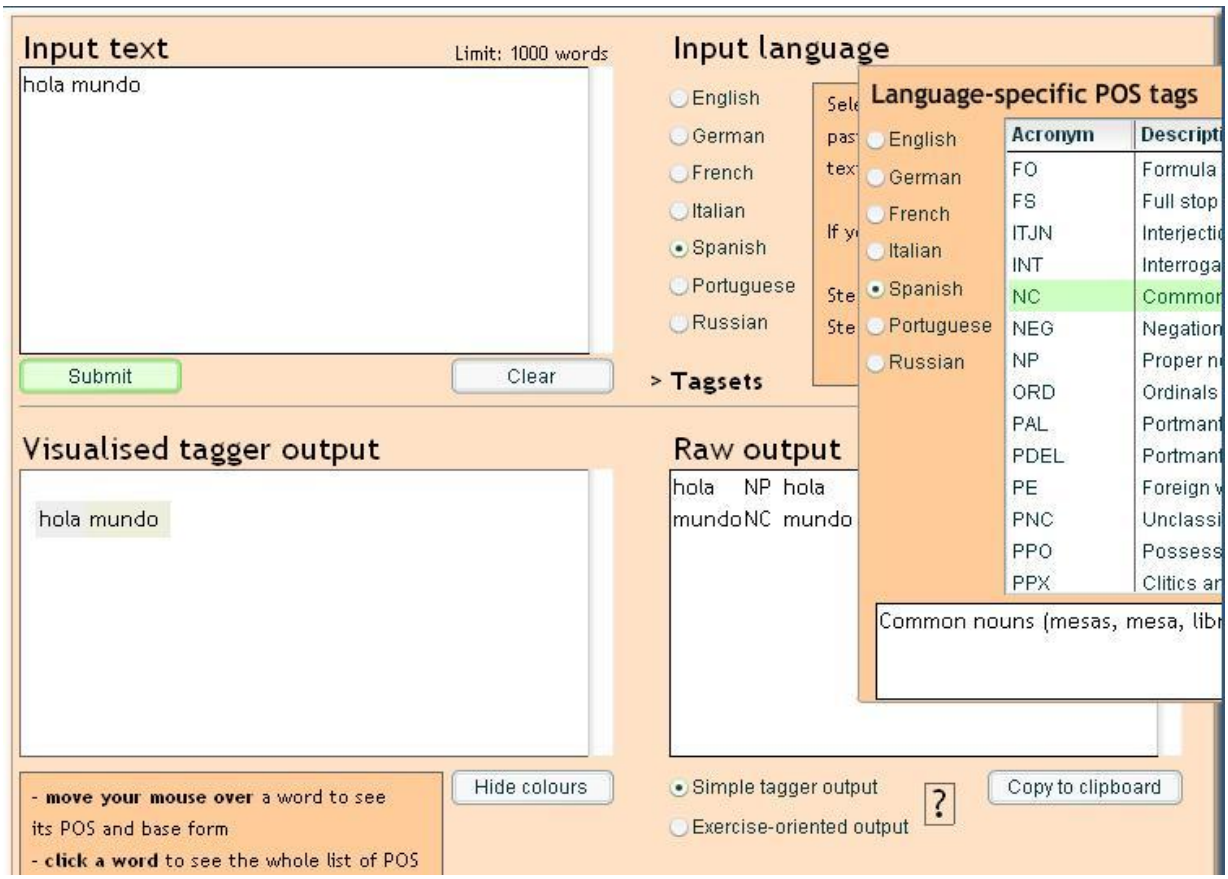

Figura 10: Ejemplo de etiquetación para el idioma español

2.3.1.3. TnT Tagger

TnT Tagger (Brants, 1998), es otro etiquetador estadístico muy eficiente, capaz de recibir entrenamiento en varios idiomas y virtualmente cualquier juego de etiquetas. TnT no está perfeccionado para ningún idioma en particular, sino que está diseñado para cubrir una gran variedad de corpus. Adaptar el etiquetado a un nuevo idioma, nuevo dominio y nuevo juego de etiquetas es muy sencillo, y además, TnT está optimizado para que sea rápido.

Esta herramienta fue desarrollada entre los años 1993 y 1999 por Thorsten Brants, profesor e investigador de la Universidad de Sarre en Alemania, bajo distribución no libre. La aplicación está

creada en ANSI[23] C usando el compilador GNU[24] C, por lo que el usuario debe estar familiarizado con los sistemas Unix y sus comandos.

La instalación en Unix o en alguna plataforma Linux es sencilla. Algunas otras plataformas pueden no estar soportadas. El programa aparece en un fichero comprimido (gzip u otro), por defecto llamado ***tnt-lin.tar.gz*** para las versiones de Linux y ***tnt-sol.tar.gz*** para las versiones de Solaris. Este fichero debe ser movido a un directorio determinado por el usuario, donde se creará un subdirectorio **tnt** que contendrá todos los ficheros necesarios y será el de la instalación.

Para que el fichero quede descomprimido se debe escribir lo siguiente: ***gzip -dc tnt-lin.tar.gz | tar xvf -*** **o** ***gzip -dc tnt-sol.tar.gz | tar xvf - ,*** lo que extrae los ficheros al directorio de instalación. Se deben ejecutar los 4 ficheros presentes y luego se pasa a habilitar el etiquetador para agregarle el nuevo idioma (español) ya que por defecto el TnT usa el alemán y el inglés.

Obtención de la herramienta TnT Tagger:

TnT Tagger está sujeto a un acuerdo de licencia, donde se dicta que no debe ser usado para fines comerciales, siendo gratis para cualquier otro fin que no sea de lucro. Para descargarlo se debe llenar el formulario con el acuerdo de licencia en *http://www.coli.uni-saarland.de/~thorsten/tnt/tnt-license.html* donde se pueden encontrar más detalles para su obtención.

2.3.1.4. Valoración de las herramientas propuestas en la Fase de Etiquetado

Entre las herramientas propuestas para esta fase hay algunas que presentan superioridades respecto a otras en cuanto a diferentes aspectos, por ejemplo, el idioma que procesan, distribución, tipo de aplicación, sistemas operativos en los que se pueden instalar y año de creación. Hay que destacar que estas herramientas se han escogido principalmente por la característica que tienen en común de trabajar con el idioma español.

[23] *American National Standards Institute*
[24] *GNU es un acrónimo recursivo que significa GNU No es Unix (GNU is Not Unix).*

Herramienta	Idioma	Distribución	Tipo de Aplicación	Sistema Operativo	Año de Creación
FreeLing	Español Catalán Italiano e Inglés	Libre	Desktop/Web	Linux/Unix MAC (OS)	2007
TreeTagger	Francés Español Otros	Libre	Desktop /Web	Linux, MAC (OS), Windows, y SPARC Workstations	1994
TnT Tagger	Todos los idiomas	Propietario	Desktop	Linux/Unix	1993-1999

Tabla 1: Comparación entre herramientas etiquetadoras

Según muestra la Tabla 1 el etiquetador TreeTagger es, de las tres herramientas, la que reúne la mayor cantidad de requisitos favorables al entorno de desarrollo de la UCI, ya que es una herramienta libre, orientada a los sistemas operativos Linux, MAC(OS), Windows, y SPARC Workstations, y su año de creación no es reciente, pero actualmente ya existen nuevas versiones implementadas. También hay que destacar que una de las ventajas que presenta este software es que se le pueden especificar nuevos idiomas con un previo entrenamiento, distintos a los convencionales (francés y español), y definidos por el usuario. Además, como se mencionaba, presta servicios on-line, aunque también tiene la característica de ser una aplicación desktop, lo que beneficia la propuesta que se plantea en esta investigación.

No obstante, el etiquetador FreeLing también es una herramienta libre, desarrollada recientemente y destinada a varios sistemas operativos, aspecto en el que es superado por el TreeTagger. Tiene versiones Web y de escritorio, lo que la convierte en una herramienta a tener en cuenta; aunque hasta la fecha funciona para el tratamiento específico de los idiomas: español, catalán, italiano e inglés.

Por su parte el etiquetador TnT Tagger, orientado a sistemas Linux/Unix, también tiene la particularidad que es entrenable en cuanto a los idiomas, es decir, es capaz de “aprender” nuevos lenguajes con un previo entrenamiento, esta característica la hace una herramienta muy fuerte, pero con la desventaja de ser software propietario.

2.3.2. Fase de Búsqueda de Conceptos

El objetivo principal de esta fase *(Valencia, 2005)* es encontrar expresiones lingüísticas que representen conceptos (ver acápite 1.5.2). Las asociaciones entre expresiones lingüísticas y los conceptos que representan, deben ser almacenadas en una base de datos que podría llamarse *base de conocimiento de conceptos (ver Tabla 2)*.

Expresiones Lingüísticas	**Concepto Asociado**
Médico	Médico
Galeno	Médico
Doctor	Médico
Ontología	Ontología
Web Semántica	Web Semántica

Tabla 2: Base de conocimiento de conceptos

A continuación se describe detalladamente como se lleva a cabo el proceso de búsqueda de conceptos en un texto. En esta explicación se usan ciertos términos que serán explicados en detalle.

El proceso de búsqueda debe tener en cuenta el hecho de que un término o concepto aparece dentro de un mismo dominio utilizando expresiones lingüísticas similares.

Se recomienda que este proceso tenga en cuenta que existen palabras que no contienen significado léxico, como las preposiciones, conjunciones, interjecciones, pronombre, determinantes, etc. Por esta razón se debe centrar la búsqueda en los sintagmas nominales[25] que tengan significado léxico, estos son: sustantivos, adjetivos, y adverbios; en el caso de los verbos tienen significado léxico, pero en esta fase el objetivo principal es la búsqueda de sintagmas nominales que representen conceptos dentro de una ontología.

Primero, ha de obtenerse cada palabra que esté etiquetada con la categoría gramatical de sustantivo, adjetivo y adverbio, en la frase que se esté analizando, y buscar si existen expresiones lingüísticas

[25] ***Sintagma nominal*** *(SN): es el sintagma o grupo de palabras cuyo núcleo está constituido por un sustantivo o pronombre o adjetivo sustantivado. También se le llama Frase nominal (FN) (Wikipedia, 2008).*

similares en la base de conocimiento de conceptos. Entonces, para cada expresión similar a la palabra actual, consideraremos si es una expresión *aceptable* y, si es así, debe añadirse esta expresión a la lista de posibles conceptos candidatos en la frase analizada. Si no se encuentra ninguna expresión similar y luego aceptable, en fin, ninguna opción válida en la base de conocimiento de conceptos, debe permitirse la posibilidad de definir los conceptos en la frase actual.

Dos expresiones lingüísticas distintas pueden representar el mismo concepto, para esto se incluye dentro de la base de conocimiento esta asociación, de modo que se pueda identificar el concepto asociado a la expresión particular. Por ejemplo, las expresiones "médico", "galeno", "doctor" representan al mismo concepto "médico".

Sería muy útil implementar dos funcionalidades para la clasificación de expresiones lingüísticas usando como referencia los datos guardados en la base de conocimientos de conceptos. Para esto se recomienda la implementación de las funciones ***similar*, y *aceptable***[26].

La función ***similar*** se encargará de identificar qué expresiones dentro de la base de conocimiento son similares a la palabra actual en un fragmento. En el caso más simple, se podría aplicar una función de "igualdad". Sin embargo, esta función no puede tratar expresiones compuestas por más de una palabra. Además, es más recomendable utilizar una función del tipo "esPrefijo". La función "esPrefijo" chequeará si la palabra actual es una subcadena de otra expresión lingüística o no.

La función "esPrefijo" tiene una desventaja importante: si la palabra actual es, por ejemplo, el sustantivo "juego", cualquier expresión que empiece con "juego", como "juego de baseball", "juego de naipes " son considerados como similares. Sería muy conveniente que esta función pudiera tratar con familias de palabras y otras peculiaridades del lenguaje. Por ejemplo, si la expresión "permite" existe ya en la base de conocimiento y el fragmento actual contiene la palabra "permitido", sería deseable que el sistema tuviese en cuenta que ambas palabras representan el mismo verbo (lema). Se recomienda para la realización de esta particularidad el uso más extensivo de lematizadores[27].

[26] *Funciones que pueden ser utilizadas en la Fase de Búsqueda de Conceptos para clasificar las expresiones lingüísticas.*

[27] ***Lematizador:*** *Software que permite, a partir de un texto etiquetado gramaticalmente, conocer el lema, es decir la raíz o la forma del diccionario de todas las palabras de un texto (Saffray, 2007).*

La función ***aceptable*** debe ser una extensión de la ***similar***. Como la función *similar* puede ser muy permisiva[28], la función *aceptable* se utiliza para determinar si la expresión lingüística actual y la expresión similar no son similares casualmente.

Además, esta función *aceptable* limita el número de palabras similares. Esta función debe ser diseñada con requerimientos fuertes: una expresión lingüística será *aceptable* si aparece en el fragmento actual, es decir, si una palabra es aceptable, entonces el fragmento actual ha de contener todas las palabras que aparecen en la expresión de la base de conocimiento.

La complejidad en la implementación de estas funciones puede variar, pero mientras más exacto sea el reconocimiento de términos "similares" y "aceptables", más precisos serán los conceptos en la base de conocimiento de conceptos y por tanto el conocimiento inferido.

A continuación se muestran algunas herramientas existentes para la extracción de términos de un texto.

2.3.2.1. TermExtract

TermExtract es una herramienta libre para la extracción de términos monolingüe desarrollada en el lenguaje Perl por Michael Schilli, programador y comercializador de scripts en Perl radicado en San Francisco, California, USA.

Esta herramienta presenta las siguientes características *(Thurmair)*:

- Analiza corpus que consta de uno o varios archivos. Los corpus usualmente consisten en varios archivos, y los usuarios no necesariamente deben estar familiarizados con la tarea de comparar y fusionar las listas de candidatos a términos similares.

- Analiza las palabras individuales y palabras compuestas. Esta herramienta utiliza tecnología híbrida (normalización lingüística, análisis de frecuencia, análisis de varias frases nominales)

[28] *Que acepta muchas palabras similares.*

para determinar los términos candidatos. Esto se hace para los 11 idiomas oficiales de la Unión Europea, además de algunas lenguas eslávicas (ruso, serbo-croata, etc.)

- No utiliza medios sofisticados de determinación de relevancia. Básicamente cuenta frecuencias para las formas normalizadas lingüísticamente. Sin embargo, reduce el ruido en su salida por filtrado lingüístico y normalización, como se describe, mediante la aplicación de filtros especiales en algunas partes de los términos a eliminar, por ejemplo, preposiciones que rara vez ocurren en los términos (así como, *en frente de, cerca, al suroeste de*), y por permitir a los usuarios especificar una lista de detención de los términos que no quieren ver. El objetivo es evitar que los usuarios tengan que escanear a través de lo que ellos podrían considerar como material inútil.

- Presenta los resultados en forma de una lista delimitada que pueden ser fácilmente visualizados, impresos, y editados en una herramienta estándar como Microsoft Excel. La salida contiene el término, su frecuencia, su etiquetación, y un número de usos definidos de frases de contexto, además de alguna información adicional que los usuarios pueden seleccionar.

- Permite escanear fácilmente a través de las listas de salida en Microsoft Excel, y el marcado de las palabras que no son consideradas a ser términos. Trabajando en un documento en el cual existan pocos términos frecuentes asegura que al menos un tiempo limitado de edición es usado de la manera más óptima.

Se puede descargar la versión 0.01 (liberada el 10 de febrero del 2008) y la versión 0.02 (liberada el 9 de marzo del 2008) de esta herramienta de las siguientes direcciones Web:

Versión 0.01:

http://search.cpan.org/CPAN/authors/id/M/MS/MSCHILLI/Text-TermExtract-0.01.tar.gz

Versión 0.02:

http://search.cpan.org/CPAN/authors/id/M/MS/MSCHILLI/Text-TermExtract-0.02.tar.gz

2.3.2.2. Copernic Summarizer

Copernic Summarizer es una herramienta cuyo principal objetivo es la creación automática de resúmenes a partir de un documento dado, sin embargo, se ha presentado sólo por su habilidad para efectuar también la extracción de conceptos, destacándolos dentro del resumen creado. Esta última característica, permite revisar la importancia de los términos extraídos al visualizar los mismos en los párrafos donde aparecen.

La versión de Copernic Summarizer que se refiere corresponde a la 2.0 de diciembre del 2001.

Características Generales de esta herramienta *(Copernic)*:

- Produce el documento resumen luego de procesar contenido texto, páginas Web, hipervínculos, mensajes e-mail y otros ficheros.
- Puede resumir un documento de un fichero sin abrirlo.
- Puede resumir una página Web desde un URL dado sin abrirla.
- Resume texto incluso del clipboard.
- Puede resumir documentos en inglés, francés, alemán y español.
- Los formatos de salida de los sumarios son varios, como HTML, XML, Rich Text Format, y Formato Texto común.
- Se puede obtener resúmenes y textos analizados muy rápidamente, ya que la opción para usar la herramienta se integra en aplicaciones como Microsoft Internet Explorer, Netscape Navigator, Adobe Acrobat Reader, Microsoft Outlook Express, Eudora, Microsoft Word, y Microsoft Outlook mediante un botón.

Configuración de Copernic Summarizer *(Copernic)*:

- Puede configurarse la longitud del sumario por defecto, en términos de porcentaje de palabras o en términos de cantidad de palabras.
- Puede configurarse para mostrar una sección de conceptos de más relevancia.
- Pueden configurarse extensiones especiales y características integradas.
- Pueden configurarse opciones visuales para resúmenes y conceptos.
- Pueden configurarse parámetros como ajustes de lenguaje y detección de proxis.

Requerimientos para el funcionamiento de Copernic Summarizer *(Copernic)*:

- Explorer 5.x
- Microsoft Outlook 97/98
- Netscape Navigator

Para adquirir el software se debe emitir un pago en la siguiente dirección Web: *http://www.copernic.com/en/store/buy-summarizer.html*

Al comprarlo se describe en detalle los pasos para su instalación, solo para el sistema operativo Windows.

Algunas muestras de la interfaz de esta herramienta, así como sus principales ventajas se pueden encontrar en: *http://www.copernic.com/en/products/summarizer/tour/index.html*

2.3.2.3. TermoStat

TermoStat (Drouin, 2003) es una herramienta para la extracción de terminologías desarrollada en Perl por Patrick Drouin, profesor asistente con el departamento de la lingüística y la traducción de la Universidad de Montreal, Canadá.

Originalmente esta herramienta fue diseñada para la extracción de términos en francés e inglés, pero actualmente cuenta también con versiones para el español y el italiano. Su principio básico es la consideración de que los términos están fuertemente relacionados con el área de especialidad a la que pertenecen. Sin embargo, TermoStat no explota recursos como diccionarios o memorias de traducción, sino que se basa en la comparación de un corpus especializado con uno de carácter general.

Por ello TermoStat requiere de un corpus de entrada conformado por dos subcorpus: el corpus en el que se quiere hacer la búsqueda de términos, que es de carácter técnico y se conoce como *corpus de análisis,* y un corpus no técnico, el *corpus de referencia,* que sirve para determinar qué tan estrecha es la relación de una palabra con respecto al corpus técnico. La diferencia en el léxico que se utiliza en los dos corpus es explotada para la identificación de candidatos a términos.

La determinación de la especificidad de una palabra con respecto al corpus de análisis se basa en su frecuencia de aparición en el corpus de análisis con respecto a su frecuencia de aparición en el corpus de referencia.

Además de la lista de candidatos obtenida (de una sola palabra), se conservan únicamente aquellas palabras cuya categoría gramatical sea nombre o adjetivo, ya que son las categorías más comunes de palabras con fuerte contenido semántico en los términos. Estas palabras son llamadas *pivotes lexicales especializados*.

Luego de la identificación de pivotes, es posible comenzar con la extracción final de candidatos a término. Todas las palabras del corpus que no se encuentren dentro de la lista de pivotes son consideradas las fronteras entre los posibles candidatos a término.

A continuación, se realiza la búsqueda de candidatos con base en su patrón morfosintáctico. Dichas cadenas candidatos son secuencias de sustantivo + adjetivo (para el caso del idioma español) en las que al menos uno de sus elementos se encuentran en la lista de pivotes lexicales.

Una de las ventajas que implica esta técnica sobre otras es la capacidad de extraer términos de una sola palabra, lo cual con otros métodos implica la generación de mucho ruido[29].

Esta herramienta puede ser utilizada a través de su sitio Web en: *http://olst.ling.umontreal.ca/~drouinp/termostat_web/?lang=fr_CA*

A continuación se muestra la interfaz online que tiene esta herramienta:

[29] *El **ruido** es la proporción de falsos positivos respecto al total de términos predichos (Carrera, 2007).*

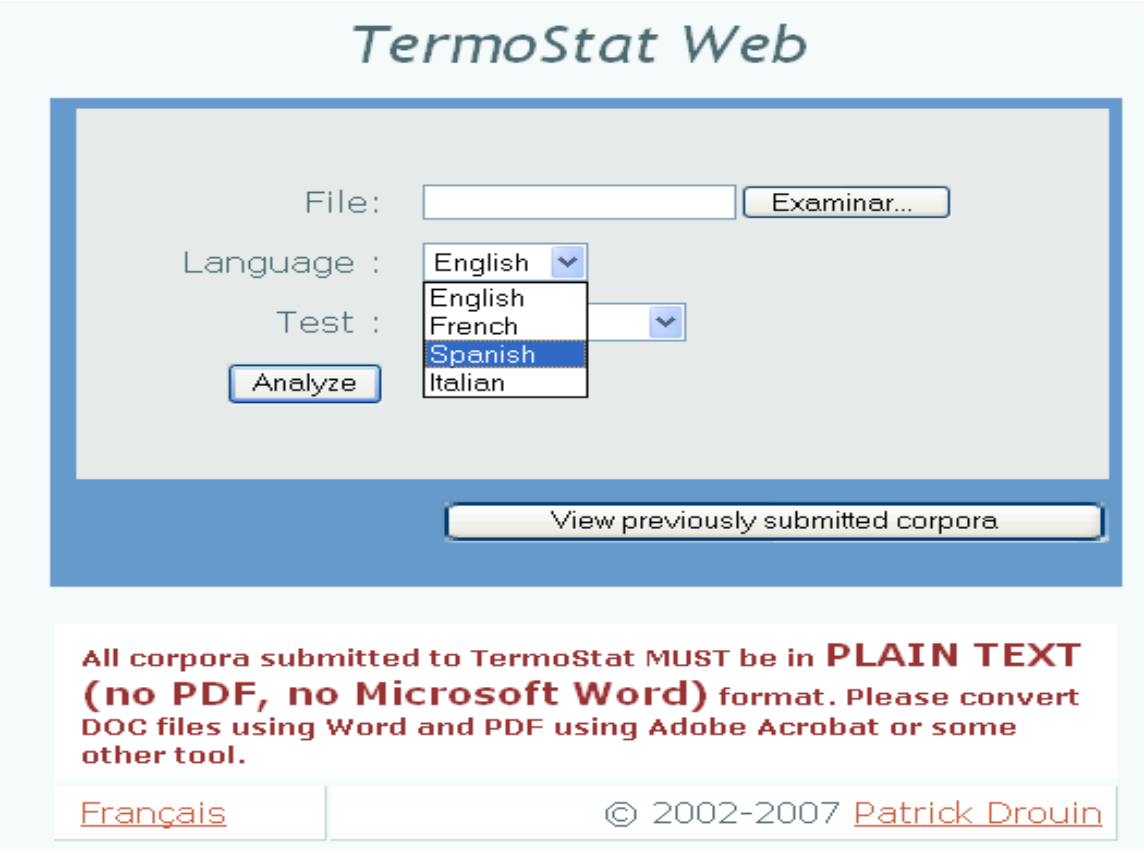

Figura 11: Interfaz online de la herramienta TermoStat (Drouin, 2003).

2.3.2.4. Valoración de las herramientas propuestas en la Fase de Búsqueda de Conceptos

En esta fase también se han propuesto varias herramientas siguiendo el propósito de que ofrezcan servicios para el idioma español. A continuación se muestra una tabla comparativa en la cual se tienen en cuenta los aspectos mencionados en el epígrafe 2.3.1.4.

Herramienta	Idioma	Distribución	Tipo de Aplicación	Sistema Operativo	Año de Creación
TermExtract	Inglés Francés Alemán Español Ruso Serbo-croata Otros	Libre	Desktop	Linux	2008
Copernic Summarizer	Inglés Francés Alemán y Español	Propietario	Desktop	Windows	2001
TermoStat	Inglés Francés Español e Italiano	Libre	Web	Todos	2003

Tabla 3: Comparación entre herramientas extractoras de términos

De las herramientas comparadas en la Tabla 3, se recomienda el extractor de términos TermExtract, para incorporar a la propuesta teórica en la fase de búsqueda de conceptos, ya que es una herramienta libre, y presta servicios a una amplia gama de idiomas, incluyendo las 11 lenguas oficiales de la Unión Europea y además, es una herramienta desktop implementada para el sistema operativo Linux que fue liberada recientemente.

Hay que mencionar también que el extractor TermoStat, es una herramienta libre que igualmente fue liberada recientemente, aunque no cuenta con un amplio tratamiento de idiomas como el software TermExtract, y los servicios que presta pueden ser utilizados desde cualquier sistema operativo, ya que es una aplicación Web.

Al contrario de las herramientas anteriores, Copernic Summarizer es software propietario, característica que lo pone en desventaja respecto a las otras aplicaciones. Como se ha mencionado en el epígrafe 2.3.2.2., el principal objetivo de Copernic Summarizer es la creación automática de resúmenes, aunque tiene la capacidad de efectuar la extracción de conceptos para el idioma español,

razón por la cual se ha incorporado dentro de la propuesta de herramientas, ya que existen muy pocas herramientas extractoras de términos que ofrezcan servicios para el tratamiento de este idioma.

2.3.3. Fase de Inferencia

En el procesamiento de lenguaje natural, las relaciones entre entidades de conocimiento, y más concretamente entre conceptos, suelen estar asociadas a verbos. Es por eso que esta fase se centra principalmente en la obtención de relaciones entre conceptos (ver acápite 1.5.3). Sin embargo, en esta fase también se puede inferir otras categorías de conocimiento como conceptos que no se hayan inferido en la fase anterior (Valencia, 2005).

Esta fase utiliza una base de conocimiento que contiene todas las expresiones lingüísticas que representan relaciones y un subsistema capaz de inferir los participantes de las relaciones. Es recomendable la implementación de este subsistema de inferencia usando el lenguaje de programación C++ u otro lenguaje orientado a objetos, debido al uso extensivo de estructuras de datos abstractas como árboles y listas. A continuación se explica el funcionamiento de algunos algoritmos de inferencia.

2.3.3.1. RDR (Ripple Down Rules)

RDR es un método basado en reglas, con una diferencia, los sistemas de reglas normalmente son una larga lista de trozos individuales de conocimiento y en la base de conocimiento manejada por RDR las reglas son almacenadas en un árbol. Cada nodo del árbol es una regla, en la versión más simple de RDR cada nodo tiene dos hijos (árbol binario). Un hijo es para el caso que las condiciones especificadas en la regla sean satisfechas, y el otro para cuando no lo son. Cada nodo también tiene una conclusión que debe ser adoptada si las condiciones son satisfechas *(Clark, 2000).*

El caso en cuestión comienza en el nodo raíz y atraviesa el árbol hasta que no tenga más hijos que visitar. Al final del viaje la conclusión es tomada de la última regla que fue satisfecha.

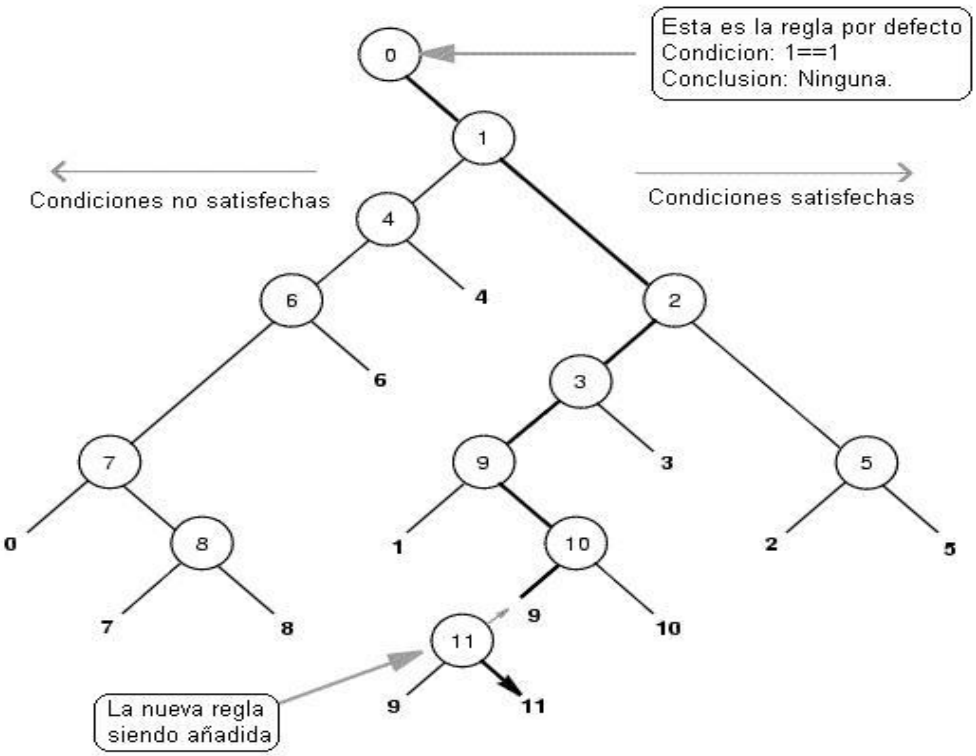

Figura 12: Estructura de una base de conocimiento RDR (Clark, 2000).

En la Figura 12 se observa una base de conocimiento en el proceso de perfeccionamiento. Las líneas más gruesas muestran el camino del caso en cuestión. Las condiciones de la regla 1 fueron satisfechas. Luego la próxima regla a ser aplicada es la regla 2 que no fue satisfecha, por lo que el caso fue a la regla 3 y así sucesivamente.

La última regla en satisfacerse fue la regla 9, entonces la conclusión de la regla 9 fue dada. Resulta que ésta no fue la conclusión deseada, entonces debe ser añadida una regla extra, que será la regla 11. El caso en cuestión satisface la condición de la regla 11 que es la nueva, así que estará enlazado con la conclusión que escribimos en la regla 11.

La belleza de este simple proceso es que no se necesita analizar el sistema completo antes de añadir una nueva regla. Las reglas son solo añadidas en los finales del árbol, así ellas solo afectarán la salida

de la regla anterior. En la Figura 12*,* la regla 11 afecta los casos que están clasificados por la regla 9, por lo que se necesita saber la diferencia entre estas dos reglas.

Para saber la diferencia entre dos reglas es más fácil trabajar desde los casos que ellas clasifican. Parte del sistema RDR es una BD de "casos angulares". Cada vez que es hecho un nuevo perfeccionamiento, el caso en cuestión es añadido a los casos angulares para futura referencia. En el caso de la Figura 12 el caso angular para la regla 9 será comparado con el caso en cuestión. La diferencia será presentada y está en manos del experto seleccionar cuáles diferencias serán usadas en la confección de la regla. A este proceso se le llama selección de diferencias, donde el operador ni siquiera necesita ver las reglas actuales, solo los casos *(Clark, 2000).*

La diferencia con otros sistemas es que aquí solo un caso angular es relevante a la adición de una regla. En otros sistemas todos los casos angulares deben ser tenidos en cuenta para chequear que no existan efectos no deseados. Queda en manos del operador actual decidir qué es importante en el sistema, por lo que es bueno tener solo una persona desarrollando esa actividad, con lo que evitaríamos la inconsistencia de datos *(Clark, 2000).*

Hay varias estructuras de tipo RDR, por ejemplo: Single Classification Ripple Down Rules (SCRDR), Multiple Classification Ripple Down Rules (MCRDR), Nested Ripple Down Rules (NRDR) y Repeat Inference Multiple Classification Ripple Down Rules (RIMCRDR).

2.3.3.2. MCRDR

La filosofía MCRDR se ha desarrollado para manejar problemas de clasificación múltiple. Su objetivo es conservar las ventajas y la estrategia esencial de RDR pero manejando múltiples clasificaciones independientes *(Valencia, 2005).*

MCRDR evalúa todas las reglas en el primer nivel de la base de conocimiento. Después, evalúa las reglas en el siguiente nivel de refinamiento para cada regla que se satisfizo en el nivel superior, y así en adelante. El proceso se detiene cuando no hay más hijos para evaluar o cuando ninguna de estas reglas puede ser satisfecha por el caso actual. De este modo, se obtienen múltiples trazos de reglas

(en los que cada trazo representa una secuencia de refinamiento particular) y, por lo tanto, múltiples conclusiones *(Valencia, 2005).*

La herramienta que ponga en práctica esta propuesta utilizando MCRDR como motor de inferencia debe obtener las palabras etiquetadas como verbos en la frase actual, y buscar expresiones lingüísticas dentro de la base de conocimiento de relaciones, para obtener el tipo de relación asociada a la expresión seleccionada (una expresión verbal en la frase actual). Esta búsqueda en la *base de conocimiento de relaciones (ver Tabla 4)* utiliza los criterios *similar* y *aceptable* definidos en la fase anterior (Valencia, 2005).

Expresiones Lingüísticas	**Relación Asociada**
es un	ES-UN
es un tipo de	ES-UN

Tabla 4: Base de conocimiento de relaciones

Obviamente, puede haber casos en los que no se encuentre ninguna buena opción para determinar una expresión lingüística que represente una relación. En este caso, el usuario debe definir el conocimiento asociado a la expresión actual.

Una vez encontrada la relación asociada a la expresión verbal principal en la frase actual, el subsistema de inferencia es utilizado para adquirir conocimiento por medio de las categorías de las palabras y su posición en la frase actual con respecto a la expresión verbal, y la relación asociada a un verbo, si existe.

Debe crearse, entonces, un caso formado por la relación, que representará la expresión verbal, y la categoría de las otras palabras de la frase. Por ejemplo, en la frase siguiente “El timón es una parte del automóvil”, si el subsistema de inferencia identificase que la expresión verbal “es una parte del” representa a una relación PARTE-DE buscando en la base de conocimiento de relaciones, entonces debe crearse un caso, como se muestra en la Tabla 5, en el que aparecerá la relación y las categorías de las palabras en la frase. Categoría$_i$ representa a la categoría de la palabra que está en la posición *i* con respecto a la expresión verbal. Así, “El” está en la posición -2, “timón” en la -1 y “automóvil” en la posición “+1”.

Fase:	El timón **es una parte del** automóvil
Expresión verbal:	es una parte del
Caso generado:	relación: PARTE-DE categoría$_{-1}$: Sustantivo categoría$_{-2}$: Determinante categoría$_{+1}$: Sustantivo

Tabla 5: Un ejemplo de un caso generado por el subsistema de inferencia

El caso creado se introduce en el sistema a implementar y dará lugar a una conclusión, la cual si es correcta, el subsistema de inferencia y su conjunto de reglas no sufre modificación alguna; pero si no, se deben obtener las diferencias entre ambos casos y crearse una nueva regla con esa información.
A continuación un ejemplo:

Se asume que el subsistema de inferencia está inicialmente vacío, si se toma el caso mostrado en la Tabla 5, no será inferido ningún conocimiento, ya que no se activarán reglas porque el subsistema de inferencia está vacío. Ahora, asumamos que el experto identifica que timón y automóvil son conceptos y que entre ellos existe una relación tipo **PARTE-DE:** timón **PARTE-DE** automóvil. En este caso se debe crear una nueva conclusión que se verá plasmada en la nueva regla que se forma, y que indica que si se encuentra ante una estructura sintáctica similar, será inferido el conocimiento.

Supongamos en la siguiente tabla que la función palabra *i* devolverá la palabra en la posición especificada, relativa a la expresión verbal. La conclusión a la que el sistema debe llegar en el presente ejemplo y que se establece en la tabla, es que se infieren dos nuevos conceptos, dados por las palabras situadas en las posiciones -1 y +1, en este caso *brazo* y *cuerpo*, y que se formará también una relación **PARTE-DE** entre estos conceptos.

Frase:	El timón **es una parte del** automóvil.
Diferencias:	**relación**: PARTE-DE **categoría$_{-1}$**: Sustantivo categoría$_{-2}$: Determinante **categoría$_{+1}$**: Sustantivo
Conclusión:	**concepto1**: **new Concepto**(palabra-1())

	concepto2: **new Concepto**(palabra+1()) **New Relacion(PARTE-DE, concepto1, concepto2)**

Tabla 6: Diferencias y conclusión que obtiene el sistema

De este conjunto de diferencias, formadas por las categorías de palabras que forman la frase y la relación, sólo se toman las categorías de las palabras que contienen el conocimiento adquirido. En el ejemplo actual, la palabra “El” no forma parte del conocimiento adquirido, así que no debe tomarse en cuenta para la generación de reglas.

A continuación la actualización de reglas del subsistema de inferencia, más cercano al código sería:

If (relacion=PARTE-DE and categoría-1=Sustantivo and categoría+1= Sustantivo) **then**

concepto1=new Concepto (palabra-1())
concepto2=new Concepto (palabra+1())
relacion1=new Relacion (PARTE-DE, concepto1, concepto2);

Ahora el futuro sistema sería capaz de inferir expresiones del tipo siguiente:

Sustantivo *expresión verbal asociada a una relación PARTE-DE* **Sustantivo**.

Y así, a medida que se vayan estableciendo nuevas reglas en el subsistema de inferencia, éste será capaz de inferir conocimiento con mayor precisión, con lo cual se irán conformando las ontologías.

Al principio el proceso es semiautomático, pero a medida que se va entrenando e infiriendo nuevo conocimiento será menos necesaria la intervención del experto, ya que el sistema cometerá menos errores y será más exacto en cada inferencia, es por esto que el proceso es casi automático.

2.4. Conclusiones del Capítulo

En este capítulo se han explicado una serie de pasos a seguir para el aprendizaje de ontologías basado en inferencia de conocimiento mediante el uso de algoritmos de inferencia como RDR y MCRDR.

La propuesta identificada se compone de tres fases secuenciales:

- Fase de Etiquetado: En esta fase se obtiene la categoría gramatical de cada palabra en la frase actual. Para ello, se sugieren varios etiquetadores descritos en el presente capítulo, donde se muestran las características de cada uno de ellos, para que los futuros programadores puedan escoger según las particularidades del sistema generador de ontologías que vayan a implementar, el tagger que más se ajuste al software en desarrollo.

- Fase de Búsqueda de Conceptos: Con la realización de esta fase se identifican las expresiones lingüísticas que representan conceptos del dominio. La asociación entre las expresiones lingüísticas y los conceptos asociados se almacenan en una base de conocimiento de conceptos. Se proponen varios extractores de términos pero aun así se recomienda que se haga extensiva la búsqueda para que se utilice el extractor más ajustado a los requisitos de la herramienta que se implementará.

- Fase de Inferencia: El conjunto de actividades comprendidas en esta fase se basa en la hipótesis de que en lenguaje natural, las relaciones entre conceptos (nivel semántico) suelen estar expresadas mediante verbos. Esta fase sirve principalmente para obtener relaciones entre términos. Aquí se utiliza una base de conocimiento que contiene las expresiones lingüísticas que representan las relaciones genéricas entre conceptos, y se explica el funcionamiento de dos algoritmos de inferencia basado en reglas que obtienen los participantes en esas relaciones.

En cada una de las fases se emiten criterios sobre las diferentes herramientas posibles a emplear en la propuesta teórica identificada en este capítulo.

Capítulo 3: Análisis de Ejemplos

3.1. Introducción

En este capítulo se muestra cómo debe funcionar un sistema que haga uso de la propuesta identificada en este trabajo, para generar automática y semiautomáticamente ontologías, a través de una serie de ejemplos de diferentes dominios.

3.2. Ejemplo 1. Dominio: Componentes de Ordenador

A continuación, se describe un ejemplo simple del dominio: Componentes de Ordenador *(Valencia, 2005)*, el cual muestra cómo sería el funcionamiento ideal de un sistema basado en la propuesta identificada en este trabajo. Inicialmente, se supone que las bases de conocimiento están vacías. Para empezar se asumirá que el sistema obtiene la siguiente frase: “Una impresora es un periférico”.

Etiquetación

El sistema primero deberá analizar la frase con el etiquetador seleccionado para obtener las categorías gramaticales de cada palabra dentro de la frase. El resultado del analizador debe ser el siguiente:

Una (Determinante) impresora (Sustantivo) es (Verbo) un (Determinante) periférico (Sustantivo).

Fase de búsqueda de conceptos

Como la base de conocimiento de conceptos estará vacía, el sistema no inferirá ningún concepto.

Fase de inferencia (basada en MCRDR)

Como la base de conocimiento de relaciones y el subsistema de inferencia estarán inicialmente vacíos, el sistema no inferirá ningún conocimiento.

Como no se habrá obtenido ningún conocimiento de la frase, el sistema deberá permitir al experto que identifique el conocimiento existente en la frase. Supongamos que el experto identifique que la expresión “es un” está asociada a una relación “ES-UN” entre los conceptos “impresora” y “periférico”.

A continuación, el sistema deberá almacenar que la expresión “impresora” está asociada al concepto “impresora” y que la expresión periférico está asociada al concepto “periférico” en la base de conocimiento de conceptos, para que la próxima vez que el sistema se encuentre con estas expresiones, o unas parecidas, pueda identificar que representan estos conceptos. En la Tabla 7 se muestra cómo quedaría la base de conocimiento de conceptos.

Expresiones Lingüísticas	**Concepto Asociado**
Impresora	Impresora
Periférico	Periférico

Tabla 7: Base de conocimiento de conceptos 1 del dominio: componentes de ordenador *(Valencia, 2005).*

Además, deberá almacenar en la base de conocimiento de relaciones que la expresión “es un” representa una relación taxonómica (*ver Tabla 8*), y por lo tanto, también deberá almacenar un nuevo caso en el sub-sistema MCRDR, que crearía la siguiente regla:

If relacion=”ES-UN” and categoria(pos(verbo)-1)=Sustantivo and categoria(pos(verbo)+1)=Sustantivo
then {
    Concepto (palabra(pos(verbo)-1))
    Concepto (palabra(pos(verbo)+1))
    Relacion (ES-UN, palabra(pos(verbo)-1), palabra(pos(verbo)+1)
}

Aquí “*pos(verbo)*” es una función que devuelve la posición de la expresión lingüística que representa la relación en la frase actual. Esta función se utiliza para obtener las posiciones relativas de las palabras con respecto a la posición de la expresión lingüística. La función “*categoria (i)*” devuelve la categoría

de la palabra que se encuentra en la posición i y la función "*palabra (i)*" devuelve la palabra de dicha posición.

Expresiones Lingüísticas	**Relación Asociada**
es un	ES-UN

Tabla 8: Base de conocimiento de relaciones 1 del dominio: componentes de ordenador *(Valencia, 2005).*

Supongamos que ahora el sistema obtiene la frase siguiente: "La Epson es un tipo de impresora".

Etiquetación

La (Determinante) Epson (Sustantivo) es (Verbo) un (Determinante) tipo (Sustantivo) de (Preposición) impresora (Sustantivo)".

Fase de búsqueda de conceptos

Como la base de conocimiento de conceptos ya no está vacía, el sistema buscaría si dentro de la frase hay conceptos que pertenezcan a esta base de conocimiento (*ver Tabla 7*). El sistema entonces inferiría que la expresión "impresora" representa al concepto *impresora*.

Fase de inferencia (basada en MCRDR)

Dada esta situación, en esta fase el sistema deberá buscar por expresiones que representen alguna relación, adquiriendo así las expresiones lingüísticas **similares** al verbo principal de la frase, "es", dentro de la base de conocimiento de relaciones (*ver Tabla 8*). Entonces el sistema deberá obtener que la expresión "es un", que representa una relación ES-UN, es similar a la expresión "es". Ahora el sistema deberá adquirir alguna expresión **aceptable** entre todas las similares (si la hay). En este caso se puede distinguir que la expresión "es un" es una expresión **aceptable** para la frase actual.

Luego el sistema podrá inferir que la expresión "es un" representa una relación ES-UN, y en el sistema de reglas se activará la regla introducida anteriormente, infiriendo "Epson ***ES-UN*** Tipo".

Como se puede ver, el conocimiento que sería inferido no es correcto. Supongamos entonces que el experto realiza la corrección: "Epson ***ES-UN*** impresora", identificando que la expresión "es un tipo de" representa la relación "ES-UN".

Entonces el sistema deberá almacenar las expresiones que representan los conceptos en esta nueva frase en la base de conocimiento de conceptos. La base de conocimiento de conceptos quedaría según muestra la Tabla 9.

Expresiones Lingüísticas	**Concepto Asociado**
Epson	Epson
Impresora	Impresora
Periférico	Periférico

Tabla 9: Base de conocimiento de conceptos 2 del dominio: componentes de ordenador *(Valencia, 2005).*

También deberá almacenar en la base de conocimiento de relaciones que la expresión "es un tipo de" representa una relación taxonómica (*ver Tabla 10*).

Expresiones lingüísticas	**Relación asociada**
es un	ES-UN
es un tipo de	ES-UN

Tabla 10: Base de conocimiento de relaciones 2 del dominio: componentes de ordenador

Ahora el sistema deberá adquirir el caso generado para el sistema de inferencia para luego obtener las diferencias entre los dos casos. Como se puede notar, los dos casos son conceptualmente el mismo caso. Entonces, al sistema no le corresponderá actualizar las reglas en el sub-sistema MCRDR, por lo que el subsistema MCRDR seguirá teniendo la misma regla.

En la Figura 13 se muestra la ontología que debería ser obtenida en el procesamiento de estas frases simples.

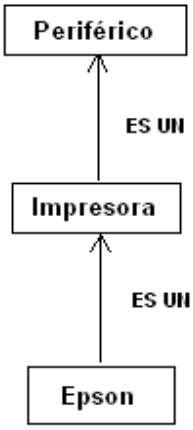

Figura 13: Ontología parcial que sería obtenida por el procesamiento del ejemplo 1 (Valencia, 2005).

Luego el sistema podrá inferir correctamente el conocimiento contenido en expresiones como "Sustantivo" es un "Sustantivo" y "Sustantivo" es un tipo de "Sustantivo".

Obsérvese que si el subsistema MCRDR contiene bastantes casos y la base de conocimiento de relaciones contiene bastantes entradas, el sistema sería capaz de inferir muchas relaciones y otras entidades de conocimiento con sólo conocer su categoría gramatical.

3.3. Ejemplo 2. Dominio: Algas Marinas

El ejemplo que se muestra a continuación presenta frases relacionadas con el dominio: Algas Marinas (ALNICOLSA, 2008) y (Cabrera, 2005).

En este ejemplo, se asumirá que la base de conocimiento de conceptos no está vacía inicialmente y está formada por seis expresiones lingüísticas que se muestran en la Tabla 11. Además, se parte de que la base de conocimiento de relaciones y el sistema MCRDR están vacíos.

Expresiones Lingüísticas	**Concepto Asociado**
Algas marinas	Alga marina
Algas terrestres	Alga terrestre

Desarrollo	Desarrollo
Desarrollo de fertilizantes	Desarrollo de fertilizante
Resistencia	Resistencia
Resistencia de los cultivos	Resistencia de cultivo

Tabla 11: Base de conocimiento de conceptos 1 del dominio: algas marinas

Supongamos que el sistema obtiene la siguiente frase: “Las algas marinas han sido utilizadas en el desarrollo de fertilizantes para la resistencia de los cultivos a las plagas”.

Etiquetación

Las (Determinante) algas (Sustantivo) marinas (Adjetivo) han (Verbo Auxiliar) sido (Verbo Auxiliar) utilizadas (Verbo Lexical) en (Preposición) el (Determinante) desarrollo (Sustantivo) de (Preposición) fertilizantes (Sustantivo) para (Preposición) la (Determinante) resistencia (Sustantivo) de (Preposición) los (Determinante) cultivos (Sustantivo) a (Preposición) las (Determinante) plagas (Sustantivo)

Fase de búsqueda de conceptos

Debido a la existencia de palabras sin conocimiento léxico, que suelen tener las categorías gramaticales de preposición, conjunción, interjección, partícula, pronombre y determinante, el sistema sólo deberá buscar los conceptos asociados a sustantivos, adjetivos y adverbios. Con esto, el sistema de búsqueda de conceptos sería más eficiente, centrándose sólo en aquellas palabras que pueden ser candidatas a ser conceptos.

Es por eso por lo que el sistema deberá adquirir cada palabra de la frase actual con categoría gramatical de sustantivo, adjetivo o adverbio y deberá buscar por palabras **similares** en la base de conocimiento de conceptos. Entonces, para cada una de estas expresiones similares, se analiza si se considera una expresión **aceptable**, como se ha indicado en el capítulo anterior.
Así, en el presente ejemplo, la primera palabra con valor léxico en la frase es “algas”, que tiene asociada la categoría gramatical de sustantivo, por lo que al sistema le corresponderá buscar las expresiones lingüísticas **similares** en la base de conocimiento de conceptos. El resultado de las palabras similares sería el siguiente:

{algas marinas, algas terrestres}

Entonces, el sistema deberá ejecutar la función **aceptable**, para encontrar si las expresiones **similares** obtenidas cumplen con el criterio de aceptabilidad. De esas dos expresiones, se puede ver que “algas marinas” es una expresión **aceptable** en esta frase, por tanto se marca como un concepto.

La siguiente palabra a procesar es “desarrollo”, porque todas las palabras anteriores no tienen una categoría gramatical de sustantivo, adjetivo ni adverbio.

El sistema deberá buscar en la base de conocimiento de conceptos, obteniendo los siguientes resultados:

Similares: {desarrollo, desarrollo de fertilizantes}

Aceptable: {desarrollo de fertilizantes}

La siguiente palabra a analizar es “resistencia“. El sistema deberá buscar en la base de conocimiento de conceptos, obteniendo los siguientes resultados:

Similares: {resistencia, resistencia de los cultivos}

Aceptable: {resistencia de los cultivos}

La siguiente palabra es “plagas”, que no tiene ninguna palabra **similar** dentro de la base de conocimiento de conceptos.

El sistema deberá encontrar entonces los tres conceptos siguientes:

“algas marinas”,”desarrollo de fertilizantes” y “resistencia de los cultivos”.

Fase de inferencia (MCRDR)

Como la base de conocimiento de relaciones y el subsistema MCRDR están inicialmente vacíos, el sistema no podría inferir ningún conocimiento en esta fase.

Supongamos que el experto identifica que la expresión “han sido utilizadas en el” está asociada a una relación “ES-USADO-POR” entre los conceptos “algas marinas” y “desarrollo de fertilizantes”.

Como las expresiones lingüísticas que representan los conceptos extraídos ya se encuentran en la base de conocimiento del sistema, éste no deberá almacenarlas.

Lo que sí se deberá almacenar en la base de conocimiento de relaciones es que la expresión “han sido utilizadas en el” representa una relación “ES-USADO-POR” (*ver Tabla 12*).

Expresiones Lingüísticas	**Relación Asociada**
han sido utilizadas	ES-USADO-POR

Tabla 12: Base de conocimiento de relaciones 1 del dominio: algas marinas

Además, se deberá almacenar un nuevo caso en el sub-sistema MCRDR, que creará la siguiente regla:

R1:
If relacion=”ES-USADO-POR” and categoria(pos(verbo)-1)=Adjetivo and categoria(pos(verbo)-2)=Sustantivo and categoria(pos(verbo)+1)=Sustantivo and categoria(pos(verbo)+2)=Preposicion and categoria(pos(verbo)+3)=Sustantivo
then {
concepto1**=Concepto** (palabra(pos(verbo)-2) + palabra(pos(verbo)-1))
concepto2**=Concepto** (palabra(pos(verbo)+1) + palabra(pos(verbo)+2) +
palabra(pos(verbo)+3))
Relacion (ES-USADO-POR, concepto1, concepto2)
}

Aquí, el operador **+** concatena dos expresiones lingüísticas.

Supongamos que ahora el sistema se encuentra con la frase siguiente: “La rodofícea es un organismo prolífero en los mares del mundo”.

Etiquetación

La (Determinante) rodofícea (Sustantivo) es (Verbo) un (Determinante) organismo (Sustantivo) prolífero (Adjetivo) en (Preposición) los (Determinante) mares (Sustantivo) del (Determinante) mundo (Sustantivo).

Fase de búsqueda de conceptos

Como la base de conocimiento de conceptos ya no está vacía, el sistema deberá buscar si dentro de la frase hay conceptos que pertenezcan a esta base de conocimiento (*ver Tabla 11*). El sistema entonces no inferirá ningún concepto por no hallar ninguna expresión **similar** en la frase actual.

Fase de inferencia (MCRDR)

Puesto que no existe ninguna expresión que represente una relación incluida en la base de conocimiento de relaciones, el sistema no deberá lanzar ninguna regla y no inferirá ningún conocimiento.

Supongamos que el experto identifica las siguientes entidades de conocimiento en la frase actual.

Conceptos: rodofícea, organismo, organismo prolífero, mares, mundo
Relaciones: rodofícea ***ES-UN*** organismo
rodofícea ***ES-UN*** organismo prolífero

Al sistema le corresponderá almacenar en la base de conocimiento de conceptos las cinco nuevas expresiones, quedando ésta como se indica en la Tabla 13.

Expresiones Lingüísticas	**Concepto Asociado**
Algas marinas	Alga marina

Algas terrestres	Alga terrestre
Desarrollo	Desarrollo
Desarrollo de fertilizantes	Desarrollo de fertilizante
Mares	Mar
Mundo	Mundo
Organismo	Organismo
Organismo prolífero	Organismo prolífero
Resistencia	Resistencia
Resistencia de los cultivos	Resistencia de cultivo
Rodofícea	Rodofícea

Tabla 13: Base de conocimiento de conceptos 2 del dominio: algas marinas

Además, se deberá introducir en la base de conocimiento de relaciones una nueva entrada, como se puede ver en la Tabla 14.

Expresiones Lingüísticas	**Relación Asociada**
han sido utilizadas	ES-USADO-POR
es un	ES-UN

Tabla 14: Base de conocimiento de relaciones 2 del dominio: algas marinas

Y el subsistema MCRDR añadirá dos nuevos casos, creando las reglas siguientes:

R2:
If relacion=”ES-UN” and categoria(pos(verbo)-1)=Sustantivo and categoria(pos(verbo)+1)=Sustantivo **then** {

Concepto (palabra(pos(verbo)-1)),

Concepto (palabra(pos(verbo)+1)),
Relacion (ES-UN, palabra(pos(verbo)-1), palabra(pos(verbo)+1)
}

R3:

If relacion=”ES-UN” and categoria(pos(verbo)-1)=Sustantivo and categoria(pos(verbo)+1)=Sustantivo and categoria(pos(verbo)+2)=Adjetivo
then {
concepto1=**Concepto** (palabra(pos(verbo)-1))
concepto2=**Concepto** (palabra(pos(verbo)+1) + palabra(pos(verbo)+2))
Relacion (ES-UN, concepto1, concepto2)
}

Supongamos que el sistema continuará analizando el texto obteniendo ahora la siguiente frase: “La rodofícea Dasya es un alga de tallo ramificado”.

Etiquetación

La (Determinante) rodofícea (Sustantivo) Dasya (Sustantivo) es (Verbo) un (Determinante) alga (Sustantivo) de (Preposición) tallo (Sustantivo) ramificado (Adjetivo)

Fase de búsqueda de conceptos

Como la base de conocimiento de conceptos ya no está vacía, al sistema le corresponderá buscar si dentro de la frase hay conceptos que pertenezcan a esta base de conocimiento (*ver Tabla 13*). El sistema entonces deberá encontrar el concepto “rodofícea”, ya que se encontraría almacenado en la base de conocimiento y deberá hallar dos expresiones **similares** para la palabra “alga” que serían:

Similares: {algas marinas, algas terrestres}

Aceptable: ninguna

Fase de inferencia (MCRDR)

Supongamos entonces que el experto identifica la palabra “alga” como nuevo concepto.

Dada esta situación, el sistema deberá buscar en la base de conocimiento de relaciones una expresión similar al verbo “es”. Hasta ese instante, sólo hay una expresión similar a la expresión “es”, que es “es un” (*ver Tabla 14*). Ahora, el sistema deberá obtener las expresiones **aceptables** de las **similares**. Se puede observar que “es un” es una expresión aceptable en la frase actual, por lo que el sistema podrá inferir que la expresión ‘es un’ representa una relación “ES-UN”, y de esta forma se activará la regla R2 infiriendo la siguiente relación:

‘Dasya ES-UN alga’

Supongamos que ahora el experto se da cuenta que el conocimiento extraído es incorrecto y lo corrige a la siguiente relación:

‘rodofícea Dasya ES-UN alga’

Ante esta situación, el sistema deberá actualizar su base de conocimiento de conceptos (*ver Tabla 15*), mientras la base de conocimientos de relaciones no se verá alterada. Sin embargo, el subsistema MCRDR deberá obtener el caso generado y las diferencias entre estos dos casos. Como se puede observar, sólo hay una diferencia consistente en que el primer concepto está formado por dos sustantivos, no sólo por un sustantivo, por lo que el sistema deberá actualizar las reglas en el subsistema MCRDR, insertando la siguiente regla:

R4:
If relacion=”ES-UN” and categoria(pos(verbo)-1)=Sustantivo and categoria(pos(verbo)-2)=Sustantivo and categoria(pos(verbo)+1)=Sustantivo
then {
concepto1=**Concepto** (palabra(pos(verbo)-2) + palabra(pos(verbo)-1))
concepto2=**Concepto** (palabra(pos(verbo)+1))
Relacion (ES-UN, concepto1, concepto2)
}

Expresiones Lingüísticas	**Concepto Asociado**
Alga	Alga
Algas marinas	Alga marina
Algas terrestres	Alga terrestre
Desarrollo	Desarrollo
Desarrollo de fertilizantes	Desarrollo de fertilizante
Mares	Mar
Mundo	Mundo
Organismo	Organismo
Organismo prolífero	Organismo prolífero
Resistencia	Resistencia
Resistencia de los cultivos	Resistencia de cultivo
Rodofícea	Rodofícea
Rodofícea Dasya	Rodofícea Dasya

Tabla 15: Base de conocimiento de conceptos 3 del dominio: algas marinas

Después de procesar las tres frases anteriores se pueden observar los componentes ontológicos que serían obtenidos en la Figura 14.

Figura 14: Componentes ontológicos parciales obtenidos por el procesamiento del ejemplo 2

3.4. Ejemplo 3. Dominio: Educación

Supongamos que las bases de conocimiento de relaciones y el sistema de inferencia están inicialmente vacíos. Supongamos además que la base de conocimiento de conceptos está formada por los conceptos mostrados en la Tabla 16 relacionados con el dominio: Educación (Wikipedia, 2008), y que el sistema obtiene la siguiente frase: “La educación a distancia es una enseñanza caracterizada por la separación física de los maestros y los estudiantes”.

Expresiones Lingüísticas	**Concepto Asociado**
Aprendizaje	Aprendizaje
Educación	Educación
Educación a distancia	Educación a distancia
Educación gratuita	Educación gratuita
Enseñanza	Enseñanza

Tabla 16: Base de conocimiento de conceptos 1 del dominio: educación

Etiquetación

La (Determinante) educación (Sustantivo) a (Preposición) distancia (Sustantivo) es (Verbo) una (Determinante) enseñanza (Sustantivo) caracterizada (Adjetivo) por (Preposición) la (Determinante) separación (Sustantivo) física (Adjetivo) de (Preposición) los (Determinante) maestros (Sustantivo) y (Conjunción) los (Determinante) estudiantes (Sustantivo)

Fase de búsqueda de conceptos

Como el sistema sólo deberá buscar conceptos asociados a sustantivos, adjetivos, y adverbios, la primera palabra a procesar en la sentencia sería "educación", que tiene asociada la categoría gramatical de sustantivo. El sistema buscaría por expresiones similares en la base de conocimiento de conceptos. El resultado de las palabras similares sería:

Similares: {educación, educación a distancia, educación gratuita}

Luego la función aceptable obtendría que la única expresión lingüística que cumple los criterios es "Educación a distancia", así que el sistema deberá obtener que la expresión "Educación a distancia" representa al concepto "Educación a distancia". La siguiente palabra para procesar sería "enseñanza", porque todas las palabras anteriores no están etiquetadas como sustantivo, adjetivo, ni adverbio.

Similares: {enseñanza}

Aceptable: {enseñanza}

El sistema deberá proceder análogamente con el resto de palabras en la frase, obteniendo los siguientes conceptos:

Conceptos: {educación a distancia, enseñanza}

Fase de inferencia

Como la base de conocimiento de relaciones y el subsistema MCRDR están todavía vacíos, no se podrá inferir conocimiento alguno en esta fase.

Supongamos que el experto identifica las siguientes entidades de conocimiento en la frase actual.

Conceptos: {educación a distancia, enseñanza} (ya inferidos por el sistema)

Relaciones: Educación a distancia ***ES-UN*** enseñanza

Como los conceptos identificados ya se encuentran en la base de conocimiento de conceptos, ésta no sufrirá ninguna actualización.

Como la base de conocimiento de relaciones está vacía, se añadiría una nueva entrada, quedando como se muestra en la Tabla 17.

Expresiones Lingüísticas	**Relación Asociada**
es una	ES-UN

Tabla 17: Base de conocimiento de relaciones 1 del dominio: educación

Y el subsistema MCRDR añadiría un nuevo caso creando la regla siguiente:

R1:
If relacion=”ES-UN” and categoria(pos(verbo)-1)=Sustantivo and categoria(pos(verbo)-2)=Preposición and categoria(pos(verbo)-3)=Sustantivo and categoria(pos(verbo)+1)=Sustantivo
then {
concepto1=**Concepto** (palabra(pos(verbo)-3) + palabra(pos(verbo)-2)+
palabra(pos(verbo)-1)),
concepto2 = **Concepto** (palabra(pos(verbo)+1)),
Relacion (ES-UN, concepto1, concepto2)
}

Supongamos que el sistema se encuentra ahora con la siguiente frase:”La educación en Cuba se trata de un proceso completamente gratuito”.

Etiquetación

La (Determinante) educación (Sustantivo) en (Preposición) Cuba (Sustantivo) se (Pronombre) trata (Verbo Lexical**)** de (preposición) un (determinante) proceso (Sustantivo) completamente (Adverbio) gratuito (Adjetivo)

Fase de búsqueda de conceptos

El sistema deberá buscar si dentro de la frase hay conceptos que pertenezcan a esta base de conocimiento (*ver Tabla 16*). En este caso encontraría el concepto siguiente:

Conceptos: {educación}

Fase de inferencia

El sistema deberá buscar en la base de conocimiento de relaciones expresiones similares al verbo "trata", pero no encontrará ninguna expresión similar y, por lo tanto, no inferirá ningún conocimiento.

Supongamos que el experto identifica que la expresión "se trata de un" está asociada a una relación taxonómica ES-UN. Después de esto, se deberá activar la regla R1 del subsistema MCRDR obteniendo las siguientes entidades de conocimiento en la frase actual.

Conceptos: {educación en Cuba, proceso}

Relaciones: Educación en Cuba ***ES-UN*** proceso

El experto deberá aceptar el conocimiento inferido y el sistema deberá introducir el concepto inferido en la base de conocimiento de conceptos para que se pueda identificar en un futuro. Esta base de conocimiento quedaría como se muestra en la Tabla 18.

Expresiones Lingüísticas	**Concepto Asociado**
Aprendizaje	Aprendizaje
Educación	Educación

Educación a distancia	Educación a distancia
Educación en Cuba	Educación en Cuba
Educación gratuita	Educación gratuita
Enseñanza	Enseñanza
Proceso	Proceso

Tabla 18: Base de conocimiento de conceptos 2 del dominio: educación

Además, se corresponderá añadir una nueva entrada en la base de conocimiento de relaciones, quedando como muestra la Tabla 19.

Expresiones Lingüísticas	**Relación Asociada**
es una	ES-UN
se trata de un	ES-UN

Tabla 19: Base de conocimiento de relaciones 2 del dominio: educación

La siguiente frase a analizar es: “La educación gratuita produce beneficios cuantiosos para la superación profesional de los estudiantes cubanos”.

Etiquetación

La (Determinante) educación (Sustantivo) gratuita (Adjetivo) produce (Verbo Lexical) beneficios (Sustantivo) cuantiosos (Adjetivo) para (Preposición) la (Determinante) superación (Sustantivo) profesional (Adjetivo) de (Preposición) los (Determinante) estudiantes (Sustantivo) cubanos (Adjetivo)

Fase de búsqueda de conceptos

El sistema deberá buscar si dentro de la frase hay conceptos que pertenezcan a esta base de conocimiento (*ver Tabla 18*). El sistema entonces deberá encontrar el concepto siguiente:

Conceptos: {educación gratuita}

Fase de inferencia

Al sistema le corresponderá buscar en la base de conocimiento de relaciones expresiones similares al verbo “produce”, pero no encontrará ninguna expresión similar y, por lo tanto, no inferirá ningún conocimiento.

El experto asociaría a la expresión “produce” una relación CAUSA e identificaría el siguiente conocimiento en la frase actual:

Conceptos: {educación gratuita, beneficios cuantiosos}

Relaciones: Educación gratuita ***CAUSA*** Beneficios cuantiosos

Luego se deberán actualizar las bases de conocimiento de conceptos y de relaciones, las cuales quedarían como muestran las Tablas 20 y 21.

Expresiones Lingüísticas	**Concepto Asociado**
Aprendizaje	Aprendizaje
Beneficios cuantiosos	Beneficios cuantiosos
Educación	Educación
Educación a distancia	Educación a distancia
Educación en Cuba	Educación en Cuba
Educación gratuita	Educación gratuita
Enseñanza	Enseñanza
Proceso	Proceso

Tabla 20: Base de conocimiento de conceptos 3 del dominio: educación

Expresiones Lingüísticas	**Relación Asociada**
es una	ES-UN
produce	CAUSA
se trata de un	ES-UN

Tabla 21: Base de conocimiento de relaciones 3 del dominio: educación

Además, se crearía la siguiente regla en el subsistema MCRDR:

R2:
If relacion=”CAUSA” and categoria(pos(verbo)-1)=Adjetivo and categoria(pos(verbo)-2)=Sustantivo and categoria(pos(verbo)+1)=Sustantivo and categoria(pos(verbo)+2)=Adjetivo
then {
concepto1=**Concepto** (palabra(pos(verbo)-2) + palabra(pos(verbo)-1))
concepto2=**Concepto** (palabra(pos(verbo)+1) + palabra(pos(verbo)+2))
Relacion (CAUSA, concepto1, concepto2)
}

La siguiente frase a procesar es: “La educación a distancia es una clase de aprendizaje que combina la educación y la tecnología”.

Etiquetación

La (Determinante) educación (Sustantivo) a (Preposición) distancia (Sustantivo) es (Verbo) una (Determinante) clase (Sustantivo) de (Preposición) aprendizaje (Sustantivo) que (Pronombre Relativo) combina (Verbo Lexical) la (Determinante) educación (Sustantivo) y (Conjunción) la (Determinante) tecnología (Sustantivo)

Fase de búsqueda de conceptos

Conceptos encontrados: {educación a distancia, educación}

Fase de inferencia

El sistema deberá buscar expresiones similares al verbo “es” en la base de conocimiento de relaciones. De esta forma obtendría que la expresión “es una” es una expresión similar y aceptable y se activaría la regla R1, obteniendo el siguiente conocimiento:

Conceptos: {educación a distancia, clase}

Relaciones: Educación a distancia ***ES-UN*** clase

Como se puede ver, este conocimiento es erróneo, así que el experto, al darse cuenta del error deberá identificar que la expresión “es una clase de” representa la relación **ES-UN**. Ahora, el sistema volvería a activar la regla R1, infiriendo el siguiente conocimiento ya correcto.

Conceptos: {educación a distancia, aprendizaje}

Relaciones: Educación a distancia ***ES-UN*** aprendizaje

El sistema ahora sólo actualizaría la base de conocimiento de relaciones (*ver Tabla 22*), ya que los conceptos identificados en esta frase ya se encontrarían en su respectiva base de conocimiento y el subsistema MCRDR no cambiaría, al haberse aplicado bien la regla y haber inferido correctamente las entidades de conocimiento.

Expresiones Lingüísticas	**Relación Asociada**
es una	ES-UN
es una clase de	ES-UN
produce	CAUSA
se trata de un	ES-UN

Tabla 22: Base de conocimiento de relaciones 4 del dominio: educación

Siguiente frase: “La educación bajo privatización es la causa principal de personas con analfabetismo“.

Etiquetación

La (Determinante) educación (Sustantivo) bajo (Preposición) privatización (Sustantivo) es (Verbo) la (Determinante) causa (Sustantivo) principal (Adjetivo) de (Preposición) personas (Sustantivo) con (Preposición) analfabetismo (Sustantivo)

Fase de búsqueda de conceptos

Conceptos encontrados: {educación}

Fase de inferencia

En esta fase se encuentra sólo una expresión similar al verbo “es” (“es una”), pero no es una expresión aceptable, por lo que el sistema no inferiría nada.

El experto identificaría que la expresión “es la causa principal” está asociada a una relación **CAUSA**. Luego se activaría la regla R2, obteniendo el siguiente conocimiento:

Conceptos: {privatización, personas con analfabetismo}

Relaciones: privatización ***CAUSA*** personas con analfabetismo

A continuación, el sistema deberá actualizarse introduciendo las expresiones verbales que representan conceptos, así como la que representa la relación **CAUSA** en sus respectivas bases de conocimiento. Las Tablas 23 y 24 muestran cómo quedarían las bases de conocimiento después de procesar estas frases.

Expresiones Lingüísticas	**Concepto Asociado**
Aprendizaje	Aprendizaje
Beneficios cuantiosos	Beneficios cuantiosos
Educación	Educación

Educación a distancia	Educación a distancia
Educación en Cuba	Educación en Cuba
Educación gratuita	Educación gratuita
Enseñanza	Enseñanza
Personas con analfabetismo	Personas con analfabetismo
Privatización	Privatización
Proceso	Proceso

Tabla 23: Base de conocimiento de conceptos 4 del dominio: educación

Expresiones Lingüísticas	**Relación Asociada**
es la causa principal	CAUSA
es una	ES-UN
es una clase de	ES-UN
puede producir	CAUSA
se trata de un	ES-UN

Tabla 24: Base de conocimiento de relaciones 5 del dominio: educación

En la Figura 15 se muestran los componentes ontológicos que serían obtenidos después de procesar las frases tratadas en este ejemplo.

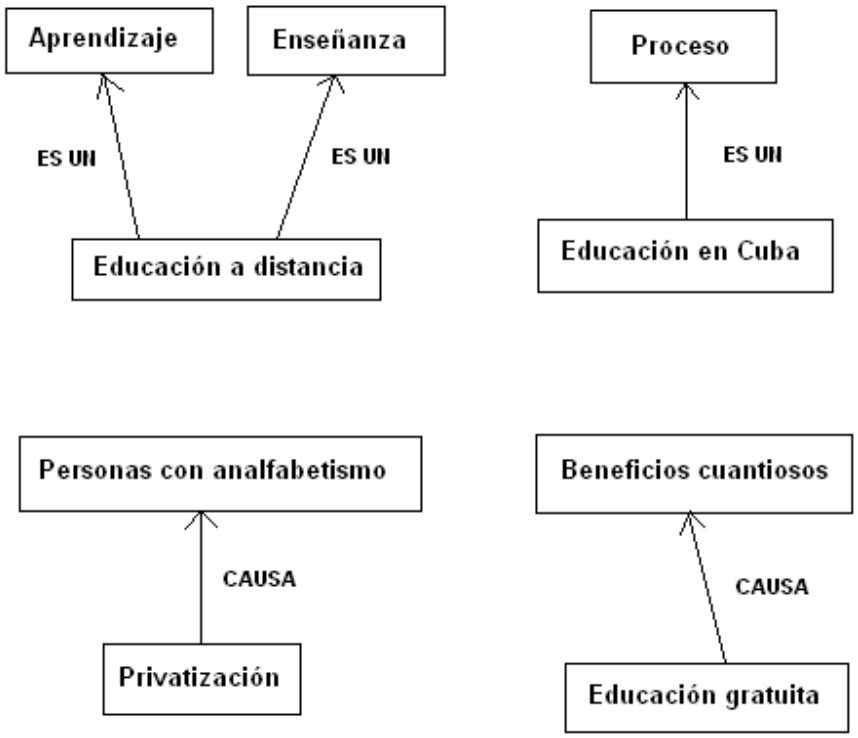

Figura 15: Componentes ontológicos obtenidos por el procesamiento del ejemplo 3

3.5. Conclusiones del Capítulo

En este capítulo se han mostrado tres ejemplos detallados que describen cómo se lleva a cabo el modus operandi de la propuesta identificada para la generación de ontologías de manera automática y semiautomática. Mediante estos ejemplos pertenecientes a diferentes dominios como: Componentes de Ordenador, Algas Marinas y Educación, se explica el proceso de adquisición del conocimiento en las tres fases en que se basa (Fase de Etiquetación, Fase de Búsqueda de Conceptos, y Fase de Inferencia).

Conclusiones

Con el cumplimiento de los objetivos trazados en el presente trabajo se ha arribado a las siguientes conclusiones:

- Se ha identificado una propuesta basada en ontologías que permite generarlas semiautomáticamente y automáticamente.

- Esta investigación sirve como guía o punto de partida para la puesta en marcha de un sistema de generación automático o semiautomático de ontologías.

- Las herramientas TreeTagger y TermExtract por las características que presentan, son las que más se ajustan a las necesidades del entorno de desarrollo de la UCI, para utilizarlas en las Fases de Etiquetado y Búsqueda de Conceptos, respectivamente.

- La aplicación del algoritmo RDR, específicamente el MCRDR en la Fase de Inferencia, lograría gran flexibilidad en la implementación de la propuesta identificada.

Recomendaciones

Para maximizar las utilidades que puede generar la aplicación de la propuesta planteada en esta investigación se recomienda:

- Aplicar la propuesta de manera flexible, ajustándola a las necesidades del entorno.
- Incorporar la propuesta obtenida al Repositorio Semántico de Objetos de Aprendizaje que se está desarrollando en la UCI, para facilitar la creación de las ontologías.
- Ir incorporando la propuesta gradualmente en los diferentes servicios informáticos de la UCI.
- Enriquecer la investigación con la búsqueda de nuevas herramientas.
- Identificar una herramienta capaz de verificar la ortografía del texto a tratar y añadirla a la propuesta planteada.
- Implementar una herramienta extractora de términos que utilice el proceso de búsqueda de conceptos expresado en la Fase de Búsqueda de Conceptos.
- Realizar una investigación sobre la integración de ontologías de diferentes dominios.

Bibliografía

Referencias Bibliográficas

(Adrián, 2007) Adrián, L. *EL TICUS. Qué son los metadatos y cómo es la organización de los recursos electrónicos. Introducción a los metadatos.* Disponible en: <http://elticus.com/?contenido=33>

(ALNICOLSA, 2008) *Las Algas.* Página Virtual de ALNICOLSA del Perú S.A.C. Disponible en: <http://taninos.tripod.com/algas.htm#usosindustriales>

(Aprendercontics) *Objeto de aprendizaje y Repositorios*. Disponible en: <http://www.espacioblog.com/aprendercontics/post/2006/08/15/objeto-aprendizaje-y-repositorios>

(Banespyme) *21 tecnologías para no tecnólogos*. Disponible en: <http://www.banespyme.org/imagesWeb/ArchivoMultimedia/Documentacion/17/serviciosweb.pdf>

(Barrón, 2007) Barrón Cedeño, Luis A. *Extracción automática de términos en contextos definitorios*, Universidad Nacional Autónoma de México, 2007. Disponible en: < http://www.iling.unam.mx/~barron/recursos/mscThesisBarronCedeno.pdf>

(Bianciotto, 2008) Bianciotto, A. *INFORMÁTICA 2.* Disponible en: < http://uniblogs.net/gregorioinformatica/>

(Borst, 1997) Borst, W.N. *Construction of Engineering Ontologies for Knowledge Sharing and Reuse.* CTIT Ph.D-thesis series No.97-14. University of Twente. Enschede, The Netherlands. (1997).

(Brants, 1998) Brants, T. *TnT -- Statistical Part-of-Speech Tagging.* Disponible en: <http://www.coli.uni-saarland.de/~thorsten/tnt/>

(Cabrera, 2005) Cabrera, R.; Clero, L.; Moreira, A.; Suárez, Ana M. *Adiciones a las algas marinas de Cuba.* Disponible en: <http://www.dict.uh.cu/Revistas/Inv_Marinas/2005/No.1/2005-009%5B1%5D.pdf>

(Cabrera & Martínez, 2007) Cabrera Pupo, K.; Martínez Gamboa, A. *Propuesta metodológica para la gestión del conocimiento basada en ontologías*. UCI, 2007

(Carrera, 2007) Carrera, S. C. *Adquisición semi-automática del conocimiento: una arquitectura preliminar*, Universidad de Vigo, 2007. p. Disponible en: < http://www.grupocole.org/cole/library/ps/Car2007a.pdf>

(Centelles, 2005) Centelles M., *Taxonomías para la categorización y la organización de la información en sitios Web*, hipertext.net, 2005.

(Cernea, 2007) Cernea, Doina A.; Del Moral, E.; Labra, E. *SOAF: un sistema de indexado semántico de OA basado en las anotaciones colaborativas*. Disponible en: <http://spdece07.ehu.es/actas/Cernea.pdf>

(Ciarán Ó Duibhín, 2008) *Windows Interface for Stuttgart Tree Tagger. Disponible en:* <http://www.smo.uhi.ac.uk/~oduibhin/oideasra/interfaces/winttinterface.htm>

(Clark, 2000) Clark, V. *To Maintain an Alarm Correlator*. Disponible en: < http://www.hermes.net.au/pvb/thesis/index.html>

(Codina, 2007) Codina, L.; Pedraza Jiménez, R.; Rovira, C. *Web semántica y ontologías en el procesamiento de la información documental.* En: *El profesional de la información,* 2007, noviembre-diciembre, v. 16, n. 6, pp. 569-578. Disponible en: < http://www.lluiscodina.com/webSemanticaOntologias2007.pdf>

(Copernic) *Copernic Summarizer 2.0*. Copernic Technologies Inc, updated in December, 2001. Disponible en: <http://www.copernic.com/en/products/summarizer/>

(DPCBD, 2006) *Diez pasos para conquistar la Biblioteca Digital*. Disponible en: <http://biblioteca.unisabana.edu.co/>

(Drouin, 2003) Drouin, P. *Term extraction using non-technical corpora as a point of leverage.* Terminology, Vol. 9, Number 1, John Benjamins Publishing Company, pp 99 -115.

(Figueroa, 2006) Figueroa, L.; Palavecino, R. *Aproximación a la diferencia entre Gestión de la Información y la Gestión del Conocimiento.* Disponible en: <http://www.cibersociedad.net/congres2006/gts/comunicacio.php?id=618&llengua=es>

(Fillottrani, 2007) Fillottrani, P. R. *Fundamento de la Web Semántica. Ontologías*, 2007. 4. Disponible en: < http://cs.uns.edu.ar/~prf/teaching/FSW07/clase7.pdf>

(FreeLing Home Page) *FreeLing 2.0. An Open Source Suite of Language Analyzers.* 2007. Disponible en: <http://garraf.epsevg.upc.es/freeling/index.php?option=com_content&task=view&id=13&Itemid=42>

(Galeana) Galeana L. *Objetos de Aprendizaje.* CEUPROMED. Disponible en: <http://www.cudi.edu.mx/primavera_2004/presentaciones/Lourdes_Galeana.pdf>

(García, 2005) García Aretio, L. *Objetos de aprendizaje. Características y repositorios*. Disponible en: <http://www.uned.es/catedraunesco-ead/editorial/p7-4-2005.pdf>

(Gómez) Gómez Pérez, A. Descripción de los cursos en el primer período. Disponible en: < http://www.dia.fi.upm.es/doctorado/challenges_esp/asignaturas/web_semantica.htm >

(Gómez, 2003) Gómez Pérez, A.; Manzano Macho, D. *OntoWeb. A survey of ontology learning methods and techniques*, 2003: 86. Disponible en: <http://www.sti-innsbruck.at/fileadmin/documents/deliverables/Ontoweb/D1.5.pdf>

(Gruber, 1993) Gruber, T. R. *A translation approach to portable ontology specifications*. *Knowledge Acquisition Vol. 5:199-220*. (1993)

(Guzmán, 2006) Guzmán Luna, J.; Torres Pardo, D.; López García, Alba N. *Desarrollo de una ontología en el contexto de la Web semántica a partir de un tesauro documental tradicional.* En: Revista Interamericana de Bibliotecología. Vol. 29, No. 2 (jul.-dic. 2006); p.79-94. Disponible en: <http://eprints.rclis.org/archive/00008649/01/Desarrollo_de_una_antolog%C3%ADa.pdf>

(Hsu et al., 2001) Hsu W.L., Wu S.H., and Chen, Y.S. *Event identification based on the Information Map – InfoMap.* In symposium NLPKE of the IEEE SMC Conference, Tuckson, Arizona, USA. (2001).

(Jacquemin, 2001) *Spotting and Discovering Terms through Natural Language Processing. MIT Press*

(Kant, 2001) Kant, I. (Published) *Lectures on metaphysics - Part III Metaphysik L2.* Cambridge University Press. (2001)

(Leibniz)(Couturat, 1903) Couturat, L. Opuscules et fragments inédits de Leibniz, Paris 1903, p. 512. Disponible en : < http://www.bariloche.com.ar/filosofia/public/publitext5.htm#_ftn1>

(López, 2005) López Guzmán, C. *Los Repositorios de Objetos de Aprendizaje como soporte a un entorno e-Learning*. Tesina doctoral, Universidad de Salamanca. Disponible en: <http://www.biblioweb.dgsca.unam.mx/libros/repositorios/>

(López, 2006) López Guzmán, C.; García Peñalvo, F.; Pernías Peco, P. *Desarrollo de repositorios de objetos de aprendizaje a través de la reutilización de los metadatos de una colección digital: de Dublin Core a IMS*. Disponible en: <http://www.um.es/ead/red/M2/lopez27.pdf>

(Lovillo, 2007) Lovillo Gil, A*. Estado del arte de los lenguajes de marcación para la web semántica*. Disponible en: <http://www.lite.etsii.urjc.es/SITIAE07docs/LenguajesMarcacionWS.pdf>

(Maedche, 2001) Maedche, A.; Staab, S. *Ontology learning for the semantic web*. En: *IEEE intelligent systems*, 2001, v. 16, n. 2, pp. 72-79.

(Maedche & Staab, 2000) Maedche A. and Staab, S. (2000). *Mining ontologies from text*. In *Proceedings of EKAW-2000, Springer Lecture Notes in Artificial Intelligence (LNAI-1937)*, Juan-Les-Pins, France, 2000. Springer, 2000.

(Martín & Olsina) Martín, María de los A.; Olsina, L. *Fortalezas y Limitaciones de las Tecnologías Actuales para el Procesamiento Semántico en la Web*. Disponible en: <http://www.frcu.utn.edu.ar/deptos/depto_3/34JAIIO/34JAIIO/asis/ASIS10.pdf>

(Meinong, 1921) Meinong. A. Self-presentation - published in: Raymond Schmidt (ed.) *Die Deutsche Philosophie der Gegenwart in Selbstdarstellung - vol. I.* (1921)

(Mendoza, 2003) Mendoza, Jorge A. *e-Learning, el futuro de la educación a distancia.* Informática Milenium, S.A.de C.V. Disponible en: < http://www.informaticamilenium.com.mx/Paginas/mn/articulo78.htm>

(Mizoguchi, 2004) Mizoguchi, R. *Ontology Engineering Environments, in S. Staab and R. Studer, editors, Handbook on Ontologies*, pages 173 – 189. Springer, 2004.

(Moreno, 2000) Moreno Ortiz, A. *Diseño e implementación de un lexicón computacional para la lexicografía y traducción automática*. Disponible en: <http://elies.rediris.es/elies9/index.htm>

(Neches et al, 1991) Neches, R.; Fikes, R.E; Finin, T.; Gruber, T.R.; Senator, T.; Swartout, W.R. (1991) *Enabling technology for knowledge sharing. AI Magazine. 12(3):36-56*.

(Pedraza, 2007) Pedraza Jiménez, R. *Generación semiautomática de ontologías.* Disponible en: <http://www.iula.upf.edu/materials/070223pedraza.pdf>

(Red TTnet, 2005) *La formación sin distancia.* Estudio realizado por el Grupo de Estudio de e-Learning de la red TTnet. Disponible en: <http://www.inem.es/otras/TTnet/pdfs/LIBRO_laformacionsindistancia.pdf>

(Rodríguez, 2007) Rodríguez Echeverría, R.; Pablos Lamas, F.; Carretero Lavado, M. *Soporte de Metadatos en Moodle.* Disponible en: <http://campusvirtual.unex.es/cala/epistemowikia/index.php?title=Soporte_de_metadatos_en_Moodle>

(Saffray, 2007) Saffray François. *Procesamiento del Lenguaje Natural. Recuperación y acceso a la información.* Disponible en: < http://procesamientolenguajenatural.50webs.com/lematizadores.htm>

(Samper, 2005) Samper Zapater, José J. *Ontologías para Servicios Web Semánticos de Información de Tráfico: Descripción y Herramientas de Explotación.* Disponible en: <http: //www.tesisenxarxa.net/TESIS_UV/AVAILABLE/TDX-0628106-085805//SAMPER.pdf.>

(Sánchez, 2002) Sánchez Calas J. C. *¿Qué son los Metadatos?* Disponible en: < http://www.uportal.cl/siel/siel_docs/estandarizacion/metadatos_SIEL.pdf>

(Sánchez, 2007) Sánchez Ruenes, D. *Domain Ontology Learning from the Web.* Disponible en: <http://www.tesisenxarxa.net/TESIS_UPC/AVAILABLE/TDX-0122108-103125//01Dsr01de02.pdf>

(Srikant & Agrawal, 1995) Srikant R; Agrawal R. *Mining generalized association rules*. In Proc. Of VLDB'95, pages 407-419.

(Studer et al, 1998) Studer, R.; Benjamins, R.; Fensel, D. *Knowledge Engineering: Principles and Methods. Data and Knowledge Engineering 25(1-2):161-197*. (1998)

(Surós & Pernía, 2006) Surós Vicente, A.; Pernía Rodríguez, R. *Repositorio de Objetos de Aprendizaje para la reutilización de contenidos en plataformas de teleformación.* UCI, 2006.

(Thurmair) Thurmair, G. *Making Term Extraction Tools Usable*: 10. Disponible en: < http://www.eamt.org/archive/dublin/THURMAIR.PDF>

(TreeTagger) *TreeTagger - a language independent part-of-speech tagger.* Disponible en: <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html>

(Valencia, 2005) Valencia García, R. *Un Entorno para la Extracción Incremental de Conocimiento desde Texto en Lenguaje Natural.* Disponible en: <http://www.tesisenred.net/TESIS_UM/AVAILABLE/TDR-1123105-140512//rvalencia.pdf>

(Wikipedia, 2007) *Repositorio.* Disponible en: <http://es.wikipedia.org/wiki/Repositorio>

(Wikipedia, 2008) *Árbol de decisión*. Disponible en: <http://es.wikipedia.org/wiki/%C3%81rbol_de_decisi%C3%B3n>

(Wikipedia, 2008) *Educación a distancia.* Disponible en: <http://es.wikipedia.org/wiki/Educaci%C3%B3n_a_distancia>

(Wikipedia, 2008) *E-learning.* Disponible en: <http://es.wikipedia.org/wiki/E-learning>

(Wikipedia, 2008) *EuroWordNet.* Disponible en: <http://es.wikipedia.org/wiki/EuroWordNet>

(Wikipedia, 2008)*Framework.* Disponible en: <http://es.wikipedia.org/wiki/Framework>

(Wikipedia, 2008) *RDF Schema*. Disponible en: <http://es.wikipedia.org/wiki/RDF_Schema>

(Wikipedia, 2008) *Sintagma nominal*. Disponible en: <http://es.wikipedia.org/wiki/Sintagma_nominal>

(Wikipedia, 2008) *WordNet.* Disponible en: <http://es.wikipedia.org/wiki/WordNet>

(Wikipedia, 2008) *XML Schema*. Disponible en: <http://es.wikipedia.org/wiki/XML_Schema>

(WTNID, 2008) *Webster's Third New Internacional Dictionary*. Disponible en:<http://translate.google.com.cu/translate?hl=es&sl=en&u=http://mwu.eb.com/&sa=X&oi=translate&resnum=1&ct=result&prev=/search%3Fq%3DWebster%25E2%2580%2599s%2BThird%2BNew%2BInternacional%2BDictionary%2Bonline%26hl%3Des%26sa%3DG>

(Woodley, 2003) Woodley, Mary S. *Glosario DCMI*. Disponible en: <http://es.dublincore.org/documents/usageguide/glossary.shtml>

(Wu & Hsu, 2002) Wu S.H; Hsu W.L. *SOAT: A Semi-Automatic Domain Ontology Acquisition Tool from Chinese Corpus.* In the 19th International Conference on Computational Linguistics, Howard International House and Academia Sinica, Taipei, Taiwan.

Bibliografía Consultada

Andreu, R.; Sieber, S. *La gestión integral del conocimiento y del aprendizaje.* Disponible en <http://www.cema.edu.ar/~jm/Clase_4/Gestion_integral_del_conocimiento.doc>

Castro Rangel, B.; Sierra Garduño, Carlos M. *Portal Semántico*, 2003. Disponible en: < http://ada.fciencias.unam.mx/~csierra/expo2.html#L0>

Céspedes, Z. R. *Las ontologías como herramienta en la Gestión del Conocimiento*: 10. Disponible en: < http://www.congreso-info.cu/UserFiles/File/Info/Info2006/Ponencias/208.pdf>

Maedche A. and Staab S. (2003) Ontology Learning. In S. Staab & R. Studer (eds.) *Handbook on Ontologies in Information Systems*. Springer 2003. Disponible en: <http://www.aifb.unikarlsruhe.de/WBS/sst/Research/Publications/handbook-ontology-learning.pdf>

Chiarani, M.; Pianucci, I.; Leguizamon, G. *Repositorio de Objetos de Aprendizaje para Carreras Informáticas*: 7. Disponible en: <http://www.dirinfo.unsl.edu.ar/~profeso/PagProy/articulos/736-WICC_2006_chiarani.pdf>

Procesamiento de Lenguajes Naturales: 4. Disponible en: < http://procesamientolenguajenatural.50webs.com/pdf/Procesamiento%20de%20Lenguajes%20Naturales.pdf>

Abascal, R.; Rumpler, B. *Evaluación de Herramientas de Extracción Automática de Conceptos Dentro de un Ambiente de Biblioteca Digital*: 18. Disponible en: < http://editorial.unab.edu.co/revistas/rcc/pdfs/r61_art1_r.pdf>

Schulte Im Walde, S.; Zinsmeister, H. *Introduction to Corpus Resources, Annotation and Access: Tree Tagger*, 2006: 1. Disponible en: <http://www.coli.uni-saarland.de/~schulte/Teaching/ESSLLI-06/Exercises/tree-tagger-handout.pdf>

Glosario de Términos

1. **Árbol de decisión:** Es un modelo de predicción utilizado en el ámbito de la inteligencia artificial, dada una base de datos se construyen estos diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que suceden de forma sucesiva, para la resolución de un problema *(Wikipedia, 2008).*

2. **e-Learning:** Conjunto de tecnologías, aplicaciones y servicios orientados a facilitar la enseñanza y el aprendizaje a través de Internet/Intranet, que facilitan el acceso a la información y la comunicación con otros participantes *(Red TTnet, 2005).*

3. **Gestión del Conocimiento:** Proceso sistemático de buscar, organizar, filtrar y presentar la información con el objetivo de mejorar la comprensión de las personas en un área específica de interés *(Figueroa, 2006).*

4. **Lematizador:** Software que permite, a partir de un texto etiquetado gramaticalmente, conocer el lema, es decir la raíz o la forma del diccionario de todas las palabras de un texto *(Saffray, 2007).*

5. **Metadatos:** Información que describe datos que incluyen el contenido, la forma y las características técnicas y editoriales de la información electrónica, los cuales son generados, consultados, manipulados y distribuidos en la red *(Sánchez, 2002).*

6. **Objetos de Aprendizaje:** Cualquier recurso con una intención formativa, compuesto de uno o varios elementos digitales, descrito con metadatos, que pueda ser utilizado y reutilizado dentro de un entorno e-Learning *(López, 2005).*

7. **Ontología:** Estructura jerárquica que define formalmente las relaciones semánticas de un conjunto de conceptos. Se usa para crear vocabularios estructurados para la recuperación o el intercambio de información *(Woodley, 2003).*

8. **RDF Shema:** Es una extensión semántica de RDF. Un lenguaje primitivo de ontologías que proporciona los elementos básicos para la descripción de vocabularios. La versión más reciente fue publicada en febrero de 2004 también por la W3C *(Wikipedia, 2008).*

9. **Repositorio:** Colección de recursos digitales, accesibles a través de una red. Pueden incluir los recursos, los metadatos que describan dichos recursos o ambos *(Bianciotto, 2008)*.

10. **Servicio Web:** Es una colección de protocolos y estándares, cuya función es el intercambio de datos entre aplicaciones a través de Internet. La interacción de aplicaciones de distintos sistemas operativos permite que programas de muy diversa concepción se combinen y proporcionen servicios integrados. Ésta es la nueva Internet, centrada en las aplicaciones y basada en los servicios *(Banespyme).*

11. **Taxonomía:** Es un conjunto organizado de términos utilizado para organizar información y pensado principalmente para poder navegar por él. En otras palabras, una taxonomía simplemente exige que sus componentes estén organizados de manera que se puedan "recorrer". Su rasgo fundamental y definitorio es, por tanto, su finalidad de exploración. Son usadas, por ejemplo, para la categorización y la organización de la información en sitios Web *(Centelles, 2005).*

12. **Web Semántica:** Extensión de la Web actual que cuenta con información debidamente estructurada, lo que proporciona un significado bien definido. Mejora la forma en la que las máquinas y las personas trabajan en cooperación *(Samper, 2005).*

13. **XML Schema:** Es un lenguaje de esquema utilizado para describir la estructura y las restricciones de los contenidos de los documentos XML de una forma muy precisa, más allá de las normas sintácticas impuestas por el propio lenguaje XML. Se consigue así, una percepción del tipo de documento con un nivel alto de abstracción. Fue desarrollado por el W3C y alcanzó el nivel de recomendación en mayo de 2001 *(Wikipedia, 2008).*